The effect of category information on feature formation

Emily Gleason, Brown University
Honors Thesis
April, 2015

Advised by Professor Joseph Austerweil

In partial fulfillment of the requirements for the degree of Bachelor of the Sciences with Honors in Cognitive Science for the Department of Cognitive, Linguistic, and Psychological Sciences

Acknowledgments

I would like to thank my thesis advisor, Professor Joseph Austerweil, for all of his continued support. Professor Austerweil was a source of constant encouragement, advice, and enthusiasm, and this thesis would not have been possible without all of his help.

I would also like to thank Professors Karen Schloss and Steven Sloman for their feedback, which helped improve this research and make it more communicable.

I would like to thank the Karen T. Romer Undergraduate Research and Teaching Award for providing funding for part of this research.

Finally, I would like to thank Brown University for providing me with the opportunity to do this project as well as resources to make the research possible.

Abstract
There are many factors that influence the features we use to represent objects. Previous work has indicated that people encode parts of objects as features that are diagnostic for categorization (Murphy & Schyns, 1994). However, how categorization information is used to influence feature formation is unknown. Category information may be treated as a highly salient part of the object, but otherwise no different from visual parts (the *label* hypothesis), or as a cue for how to organize the stimuli into sets that should be encoded in a similar manner (the *structure* hypothesis). Following Austerweil and Griffiths (2013), we formalized these two hypotheses (IBP+ and IBF for the *label* and *structure* hypotheses respectively) within their rational framework and then used them to develop a novel experiment where they make diverging predictions. The experimental results provide support for the IBP+ model.

Introduction

Categorization is a critical process in cognition by which we group like objects together. It is by this process that we are able to make proper inferences about known objects and generalize knowledge onto new instances. For example, if we encounter a new animal it is extremely useful to be able to categorize this animal as dangerous or friendly. Categories are an efficient way to sort the world into meaningful and manageable pieces.

One question is how categories are mentally represented. One possible theory is that categories are defined by sets of features (Murphy, 2002). The most common way of thinking of features is as small pieces that can be put together into objects. Recognizing the features allows one to make comparisons between objects and inferences about new objects. In addition, differences among the features for each individual person may explain the different reactions individuals have to certain stimuli, such as modern art (Austerweil & Griffiths, 2013). Category features work much the same way—each category can be represented by sets of features, and objects within the category will have some combination of those features (Murphy, 2002). For example, one can think of the category TABLE as defined by the features TOP and LEGS, and each table in the category will get some combination of these features. This is more flexible and efficient than memorizing each instance of a table, and much more easily generalizable. With this structure, one can ask whether or not an object belongs to the category by asking if it has the features of the category.

The way that we determine the relevant features for a given situation is still an open question. People can define any object with an arbitrary number of features (Goodman, 1972; Murphy & Medin, 1985). In essence, the features problem is underdetermined—there are infinite possible solutions possible given the data that we have. Take a simple square. The square could be represented as four straight lines, or four corners, or perhaps one line repeated four times. In fact, it could even be represented as itself, a single square, with no component parts. The mind must choose some representation, so how does it choose among the possible representations? In addition, there must be a process for learning features because there are an infinite number of potential new objects and features that we may encounter. One only has to look at the literature on feature learning—experimenters consistently create entirely novel objects (made of entirely novel parts) for participants to examine (e.g. Schyns, Goldstone, & Thibaux 1998). It is for these reasons that feature learning must be flexible. The ideal feature set would be the fewest set of

pieces that can account for the most data. For categories, these feature sets must be able to describe the objects in the category in an efficient way. In other words, ideal features are small enough to be generalizable but large enough to be meaningful. For example, given the category "quadrilaterals" the ideal features might be four straight lines. However, the exact way that we determine the correct features is still an open question.

One proposal was that features would arise from some perceptual principles, for example, Gestalt goodness. Hoffman and Richards (1984) determined a "minima rule", that given a two dimensional outline of an object, features would form between two minima points (Figure 1). This rule seems parsimonious and logical, but constraining feature formation to this principle is not enough. For example, for the lamp in Figure 1 the minima rule would propose five features (a base, stem, switch, and two parts for the shade), but in reality most individuals would find four, combining the shade into a single part (Schyns & Murphy, 1994). Although Hoffman and Singh (1996) did demonstrate many cases where minima clearly define part boundaries for individuals, there must be additional factors taken into account during feature formation to cause the kinds of discrepancies seen in the lamp example. The following section discusses previous work that explored the way outside factors such as context and category information affect the way participants choose features for the same image.

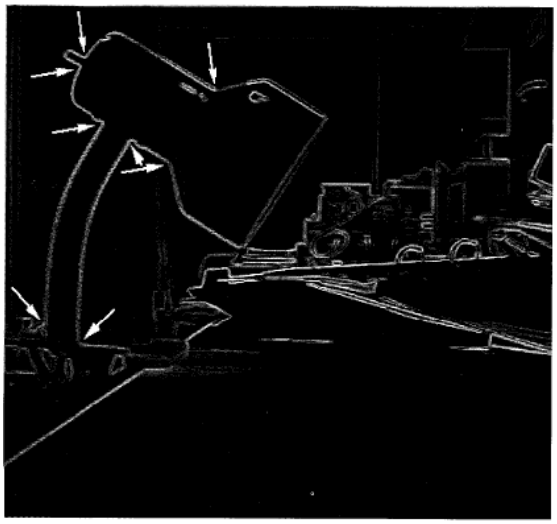


Figure 1: An example of an object outline with minima indicated by arrows reproduced from Schyns and Murphy (1994)

In this paper, we discuss previous research on factors that affect feature formation, followed by a discussion of the computational framework that we use to formalize our hypotheses and the models themselves. Finally, we present two experiments that we performed (where the second experiment controls for some concerns from Experiment 1) and discussion of all results.

Previous work

An early study by Schyns and Rodet (1997) discussed how presentation order affects feature learning. In their study, participants were presented with a series of categories of "Martian cells": X, Y, and XY. Cells in each category had the corresponding parts *x*, *y*, and *xy* respectively, where *xy* was a conjunction of parts *x* and *y*. The

experimenters manipulated the order of presentation of these parts in order to change the representation of the part *xy*. In one condition, the participants were presented with the categories in the order X, Y, XY. For this case, participants initially learned the parts *x* and *y* separately, so the part *xy* was simply a combination of known parts. In the other condition, they were presented in the order XY, X, Y. For this case, the part *xy* was learned first and was represented as a single part since participants had no knowledge of the component pieces *x* and *y*. They tested participants on the features they had learned by asking them to categorize new objects, including some of the kind X-Y, cells which had both the *x* and *y* parts not in conjunction. Participants in the X, Y, XY condition treated *xy* as a combination of parts and categorized X-Y cells as XY. Participants in the XY, X, Y condition treated *xy* as a singular feature and categorized X-Y cells as either X or Y. This provided evidence that the order in which objects are learned can lead to different feature representations. The phenomena by which two parts come to be seen as a single feature is known as unitization, and it has become a useful tool for examining situations that cause changes in feature learning (Goldstone, 1998).

Context also plays a role in feature learning, especially how pieces co-vary in the context. If two pieces are always seen together, then it would be reasonable to come to represent those pieces as a single feature. If, however, the pieces can occur separately, then they should be seen as separate features. For example, we have background knowledge that table tops are separable from table legs, so these parts should be thought of as separate. Austerweil and Griffiths (2013) demonstrated clearly that people are sensitive to this information. They presented participants with a series of objects, each with two vertical bars between two horizontal bars. In the unitized condition, the two vertical bars moved together such that they were the same distance apart in each object. In the independent condition, the two vertical bars moved independently such that they occurred at varying amounts of separation across objects. They found that only participants in the independent condition would generalize category membership to a "new separated" object, an object they had not seen before with the two bars spaced farther apart than the unitized condition. This simple experiment implies that participants are taking covariation into account when forming features.

Another major catalyst of feature learning is category diagnosticity. In other words, a part will be more likely to be learned as a feature if that part occurs in all members of the category and distinguishes members of that category from members of another category. Pevtzow and Goldstone (1994) trained participants on four objects belonging to two categories. All participants looked at the same set of objects, but there were two potential categorization schemes. For each categorization scheme, every object shared one feature with the object in the same category, and one feature with an object in the opposing category (Figure 2). The participants were then presented with whole objects followed by probes and asked whether the probe was contained within the given whole. They found that participants' reaction times were significantly faster if the given probe was a diagnostic feature than a nondiagnostic feature, implying that diagnostic features are treated differently than other parts. A strong analysis of this result would be that participants only found the features that were diagnostic, but at the very least participants certainly weighted diagnostic features much more. Clearly, category information is critical for feature learning, but the way we encode this category information is still unknown.

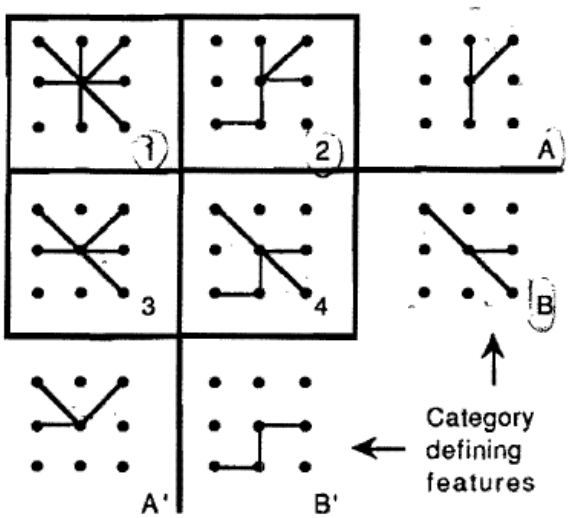


Figure 2: Figure reproduced from Pevtzow and Goldstone (1994). Objects with their associated diagnostic parts. When the horizontal line divided the objects into categories, parts A and B were diagnostic. When the vertical delineation was used, parts A' and B' were diagnostic.

In the categorization literature, there is evidence that early in life category labels are treated as a highly salient feature during the tasks of categorization and induction, but later in life "category markers" are treated differently and prioritized. Deng and Sloutsky (2012) created an experiment that pitted a highly salient feature and the category label against each other in an induction task. Both 4 to 5-yeear-old children and adults were presented with two categories of novel creatures, *flurps* and *jalets*. The novel creatures had five different components—antennae, head, arms, body, and feet. For each component the creatures could have one of two feature options. The prototypes for the two categories (never shown in training) had the opposing features for each component, and the categories were constructed such that all of the members of the category would have four shared features with category prototype. For example, all but one *flurp* had yellow star-shaped bodies, and all but one of the *jalets* had green rectangular bodies. The only feature consistent for all examples within a category were the heads. This fixed feature was made even more salient by adding motion—for *flurps* the head moved up and down, and for *jalets* the head moved side to side. Both children and adults were given an induction task where they had to choose between two different features to fit a missing component spot. The new test creatures were labeled with a category name, but for some of the objects the head feature did not match the category label. When making feature inductions for these creatures, children were more likely to follow the salient feature, the moving head, whereas adults would make inductions based on the category label. The difference in patterns of induction shows evidence for a developmental change from treating category labels as another salient feature to treating labels as true "category markers" (Deng & Sloutsky, 2012). "Category markers" are treated differently than ordinary features; they are prioritized over other features and they provide more information about the objects to which they are attached. In this study, we explore these possibilities for category labels for feature learning.

Hypotheses for how categories affect feature formation

There are two analogous hypotheses for the role of category information in feature formation. The *label* hypothesis states that category information is treated like any other sensory input (as in Anderson, 1991). In this theory, category information is appended to each object and observed similar to any other feature— category membership is another independent property of the object. The information is special only in that it is extremely salient. According to the *structure* hypothesis, each category receives its own set of feature representations taken from a shared feature repository (Austerweil & Griffiths, 2013). This means that all objects across categories contribute to creating the set of potential features, but each category has its own inference process to determine the features relevant to category determination. Because it is difficult to distinguish the predictions of these hypotheses in the feature formation domain simply by introspection, we formalized both as Bayesian models. Using these models, we were able to find certain category structures where the models make qualitatively different predictions. Given two categories, such as tables and chairs, the intuition for these differences is as follows: The *label* hypothesis predicts that features will not be shared across categories because the weight of the category information in the sensory input would be greater than the need to share information across categories. In the example of tables and chairs, this hypothesis would predict a single feature leg associated with chairs and a single feature leg associated with tables, and these features would be represented separately. Under the *structure* hypothesis, features would be shared across the categories. The leg associated with tables would likely be the same as the leg associated with chairs (Figure 3).

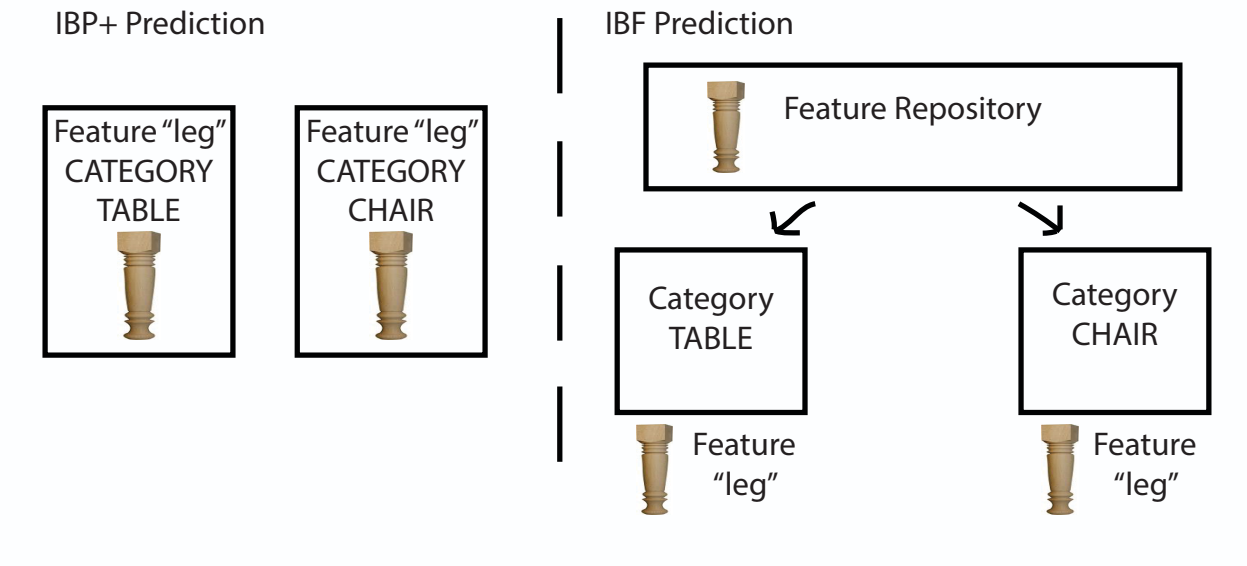


Figure 3: A visual depiction of the predictions from each hypothesis using the categories TABLE and CHAIR

Rational Analysis

Rational analysis is a technique to study cognition as the ideal solution to various problems. Because there are multiple sets of features that can work as potential solutions for each object, the feature problem can be seen as trying to find the ideal solution to an inductive problem. Using rational analysis, this solution is given by optimizing the

posterior probability of the features given any available information using Bayes' Rule (Equation 1). Bayes' Rule is used in rational analysis because it takes into account both the observed data and prior probabilities. The probability of the feature set existing independent of any data, or the prior probability, is factored into the final solution. This prior can push the final solution towards more psychologically plausible features, such as features with better Gestalt goodness or sets with fewer numbers of features overall. Factoring in the probability of the observed data being seen given that the feature set you assume is true, or the likelihood, insures that the feature set with the highest probability will accurately account for the observed data.

$$P(\text{Features}|\text{Data}) \propto P(\text{Data}|\text{Features})P(\text{Features}) \quad (1)$$

Austerweil and Griffiths (2013) modeled two ways that category information might influence feature formation: the Indian Buffet Process with Category Labels (IBP+) and the Indian Buffet Franchise (IBF). Before discussing these models, we present their computational framework for constructing feature representations.

Computational Framework

The ideal solution to the feature problem should be the set of features which best encodes all observations, enables categorization of the objects, and has relatively high *a priori* probability. The final solution can be thought of as the product of two matrices: the feature ownership matrix (**Z**) and the feature image matrix (**Y**) (Equation 2). The feature ownership matrix will indicate which features belong to each object. Each of the *N* rows of the feature ownership matrix represents an object, and each of the *K* columns represents a feature, where the cell z_{nk} will be 1 if an object *n* has feature *k* and zero if not. Each of the K rows of the feature image matrix represents a particular feature, and each column is the pixel value from a particular location for all features. The value of y_{kd} will be 1 if the pixel is on for feature *k* at location *d* and zero otherwise. The product of these matrices, X, should creates the images of observed objects (Figure 4). See Austerweil and Griffiths (2011; 2013) for more details.

$$P(\mathbf{Z}, \mathbf{Y}|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{Z}, \mathbf{Y})P(\mathbf{Z})P(\mathbf{Y}) \quad (2)$$

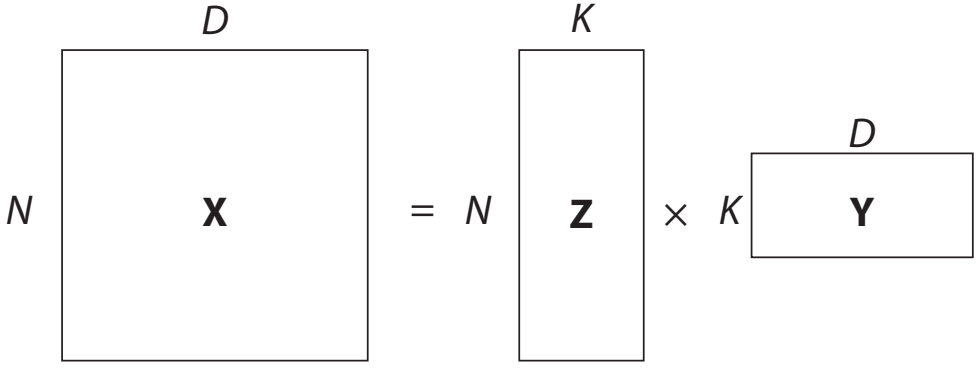


Figure 4: Feature formation as matrix decomposition.

For the prior over feature assignments, $P(\mathbf{Z})$, we use the Indian Buffet Process (IBP; Griffiths & Ghahramani, 2011) is a Bayesian nonparametric process that can be used to infer sets of feature representations without knowing the number of features *a priori*

(Austerweil & Griffiths, 2011). The process follows a simple food metaphor: an Indian buffet. Objects from a category are customers who enter an Indian buffet. All of the dishes at the buffet represent the potential features, and there are potentially infinite dishes. When a customer (object) enters, they take from dishes (features) that have already been sampled with probability relative to the number of people who have sampled the dish. Therefore, it is likely for objects to share features, since the likelihood of taking a feature increases if more objects already have that feature. In addition, the customers will sample a new dish according to a Poisson distribution with parameter α divided by the number of customers who have entered the restaurant. Because this value decreases as customers enter the restaurant (objects are observed), it becomes less and less likely for a customer to sample a new dish (for an object to have a novel feature). For most of the feature dishes the probability of being taken is essentially zero, leaving a finite number of features to represent the objects. Although the prior on the feature image matrix, Y, can be more complex to encode perceptual biases, such as a proximity bias, a simple Bernoulli prior, where each pixel is turned on independently with probability *p* (Austerweil & Griffiths, 2011) usually suffices.

The Indian Buffet Process model with category labels (IBP+)

The IBP+ formalizes the *label* hypothesis (Austerweil & Griffiths, 2013). In this model, category information is included like an additional perceptual feature, in a similar way to Anderson's (1990) Rational Model of Categorization (RMC). Category information is appended at the end of the image of each object in the matrix X as groups of ones and zeros (Figure 5). For example, category 1 would be encoded by *c* bits equal to one followed by *c* bits equal to zero (and reverse for category 2). Each feature in Y will also have information appended that indicates which category it belongs to.

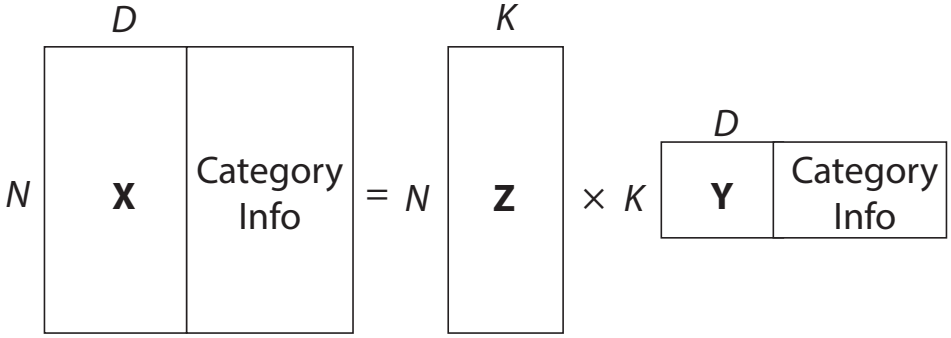


Figure 5: Matrices with appended category information.

Indian Buffet Franchise (IBF)

Austerweil and Griffiths (2013) noticed that occasionally features occur across categories and defined the IBF to capture this. The IBF formalizes the *structure* hypothesis—in this model, all objects contribute to a shared feature repository and category information is not treated as another feature. Instead, category information is included in this model in a similar manner to the hierarchical form of the RMC (Griffiths et al., 2008).

According to the IBF, each category is given its own feature ownership matrix (Equation 3). In other words, each category is able to choose its own feature distribution independently of the other categories. The potential features, however, all come from a

feature image repository shared across categories. To allow for an arbitrary number of features, since the number is still unknown, the feature images for each category are drawn from the repository generated from a Dirichlet Process (DP; Ferguson, 1973). The culinary metaphor for DP is the Chinese restaurant process (Aldous, 1985), which is directly analogous to IBP. In this case, the features are customers who arrive at a restaurant and sit at any one of an infinite number of possible tables (feature images). Customers (features) will sit at a table with other customers already seated with probability proportional to the number of customers already at the table, and at a new table proportional to a parameter β (see Austerweil & Griffiths, 2013 for more details). This way, each feature receives one image, but features from different objects and categories may have the same image because both feature image matrices are generated from the same DP (Equation 4; Figure 6 for visual representation). The chances of sitting at a table with customers increases as the customers at the table increase, but the chance of sitting at a new table decreases as customers enter the restaurant. This means there are infinite possible images for the features, but the chances of forming a new table will eventually decrease to zero, and so DP will settle on a finite feature repository.

$$\mathbf{Z}^{(\boldsymbol{\alpha})} \sim \text{IBP}\ (\alpha)\ (3)$$

$$\mathbf{Y}^{(\boldsymbol{\alpha})}|\mathbf{Y}^{(\mathbf{0})} \sim \text{DP}(\beta, \text{Bernoulli}(p))\ (4)$$

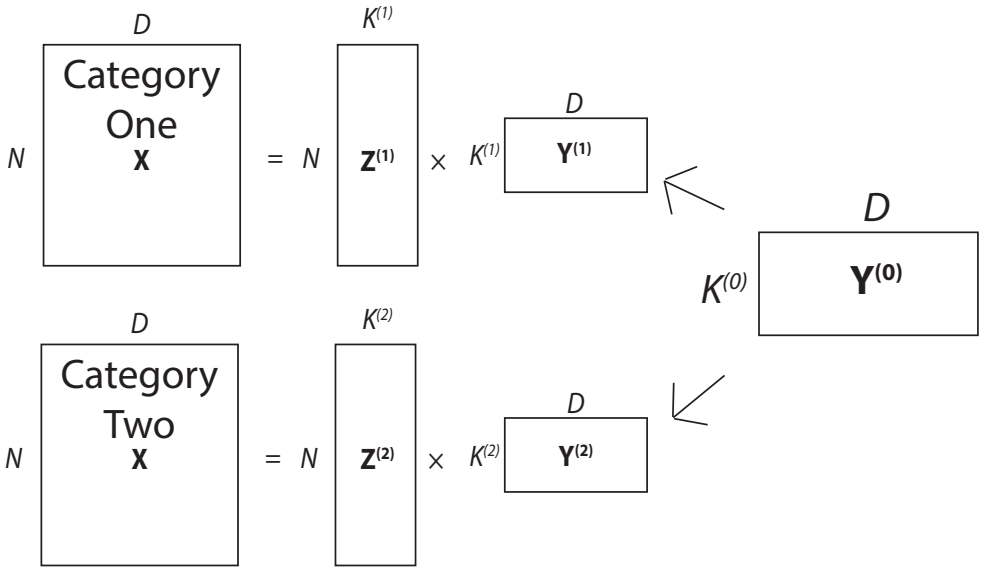


Figure 6: Feature image matrices formed from shared DP

Generalization Paradigm, example from Austerweil & Griffiths, 2011

Once a participant has learned a feature for a category, then she should believe that other objects with that feature also belong to that category. Because of this, a generalization task (asking which new objects are likely members of a learned category) is a good way to determine which features a participant has inferred.

In Experiments 1 and 2 of Austerweil and Griffiths (2011), participants were given a set of sixteen novel images to study, which they were told were inscriptions recently found by a Martian rover in a cave on Mars. There were 20 possible objects, each with three of the

six potential parts taken from a single master object (Figure 7). Participants were trained on one of two conditions. In one condition, the sixteen images they saw constituted a Unitized set. A Unitized set contained four different inscriptions repeated four times within the set. In the other condition, the sixteen images constituted a Factorial set, which had sixteen unique combinations of parts (Figure 8).

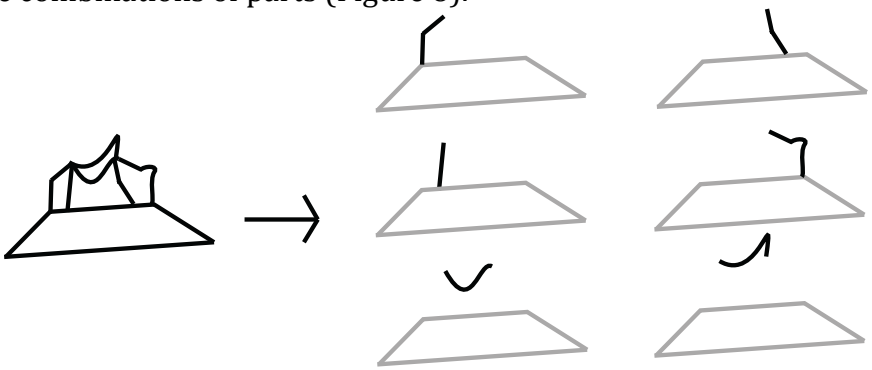


Figure 7: The master object an accompanying six independent features

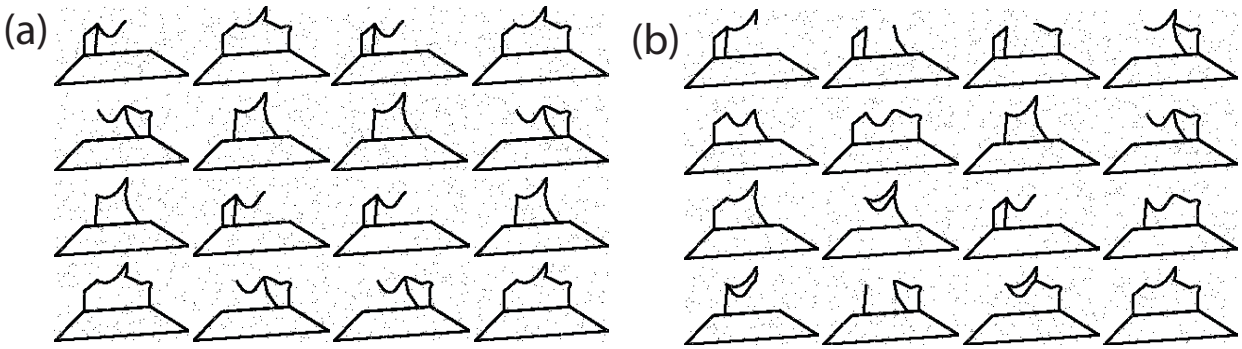


Figure 8: (a) a Unitized set and (b) a Factorial set

Participants were then asked to give likelihood ratings for several new images being found in the original cave. There were three types of test images: *seen*, *unseen*, and *shuffled*. Seen images were those they already observed during training. *Unseen* images used the same set of six potential parts but were combinations that had not been seen during training. *Shuffled* images came from the same master object, but they used "shuffled parts", parts that were broken up in a different way than the original (Figure 9).

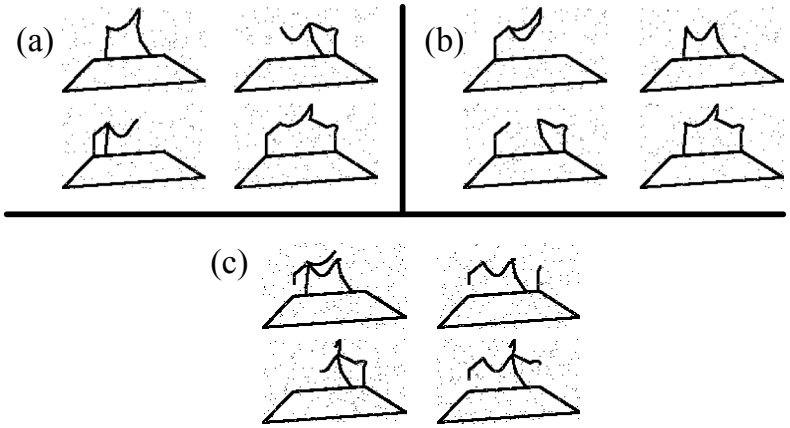


Figure 9: Examples for (a) *seen*, (b) *unseen*, and (c) *shuffled* test images

In the Unitized condition, only seen objects were given high likelihood ratings. In the Factorial condition, *seen* and *unseen* objects were both given high likelihood ratings, while the *shuffled* objects were given low ratings. Because participants should only rate objects as likely to belong to a category if they have the feature of that category, this indicates that participants in the Unitized condition used only the four given objects as features, whereas in the Factorial condition participants treated the six independent underlying parts as features. This result was replicated with the IBP model. I adapted this paradigm for my experiment.

Testing how categories affect feature formation

Austerweil and Grifftihs' (2011) experiments only looked at the effect of using one category, but what happens when two separate unitized sets formed from the same six parts are given as two categories to be analyzed at once? When these images are given to the IBP+, it infers unitized objects as the features for each category. Although this result is not perfectly stable, most results will have partial combinations of features, and the result with the highest probability is eight unitized features (c = 2200, ε = .001, λ = .999, p = .4, α = 2). Therefore, the label hypothesis predicts that the objects are learned as features, and that people will not generalize to the *unseen* objects. Because the model allows some variability in the accuracy of the final representation, the final solution is commonly six unitized features and two independent features. Critically, there was always the presence of large unitized features tied to a particular category. This makes the predictions of the IBP+ distinguishable from the IBF predictions. When the two unitized sets are given to the IBF, it infers the six independent parts as features for both categories. (ε = .01, λ = .99, p = .08, α_0= 2, α_1 = 2). Because all objects contribute to the feature repository, six features is the more efficient solution to use rather than eight. Thus, the structure hypothesis predicts people will learn the parts as features and will generalize to the *unseen* objects.

Experiment 1

Methods

Participants:

Sixteen participants took part in this study either for course credit or for pay at a rate of $10 an hour.

Stimuli:

The stimuli from this experiment were very similar to those used by Austerweil and Griffiths (2011). In Austerweil and Griffiths (2011), the shuffled objects were made from dividing the same master object into a different set of six parts, but these shuffled parts had worse Prägnanz than the experimental parts. Therefore, the ratings on the shuffled objects might have been lowered. Because of this, a new master object was created that could be split into two possible partitions with six parts each with relatively similar Prägnanz (Figure 10). Each partition was associated with a set of 20 possible objects, each made with three parts of the six potential parts. For each experiment, if the set from Partition 1 were used during training then Partition 2 would be used as shuffled images during test and vice versa. For each partition there were three Unitized sets, A, B, and C. Unitized sets consisted of four objects where each of the six parts occurred equally often and the parts were strongly, but not perfectly, correlated. Training images were printed on cardstock and

given labels to indicate the category they were from. Unitized sets A and B were used during training and C was always used for testing.

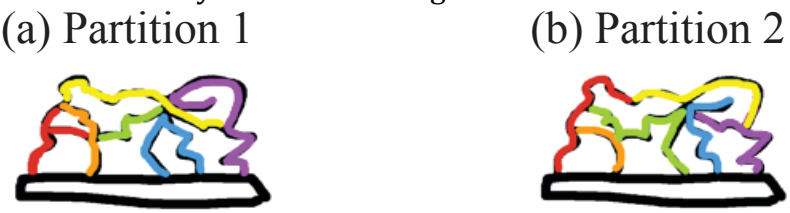


Figure 10: Parts defining (a) Partition 1 and (b) Partition 2 (colors define each part)

Design:

Eight participants were tested using Partition 1 and eight were tested on Partition 2. The two potential categories for the objects were fictional Martian caves, "CAVE DAX" and "CAVE NARL". The participants saw Unitized sets A and B from a single partition, and the order of these was counterbalanced such that A and B were in Cave Dax and Cave Narl an equal number of times across participants. There were also two potential test orders created which were counterbalanced.

Procedure:

Participants were given a set of instructions and two stacks of cards, one labeled "CAVE DAX" and the other labeled "CAVE NARL". They were told through written instructions that a rover had recently gone to Mars and found inscriptions from two different caves. They were told to study the cards for 5-10 minutes (this was self-paced).

After the study period, the participants were given a second set of instructions and a test booklet. These instructions read that new images had come in from Mars, but a competing research team had hacked the data, removed the labels, and added fake images. The participants' task was to rate on a 1-7 Likert scale how likely each image was to be real. Test images consisted of the following objects: 8 observed objects from one of the two caves (4 *seen same* and 4 *seen other*), 4 unobserved objects from the same partition (Unitized C; *unseen*), and 12 objects from the other partition (*shuffled*; see examples in Figure 11). For each object they answered four questions:

1. Could this image be found in Cave Dax?
Yes No

2. How likely is it that this is a real image from Cave Dax?
[scale from 1-7, 1 is Very Unlikely, 7 is Very Likely]

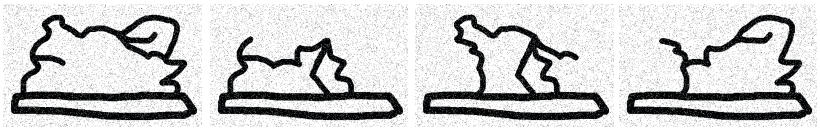
3. Could this image be found in Cave Narl?
Yes No

4. How likely is it that this is a real image from Cave Narl?
[scale from 1-7, 1 is Very Unlikely, 7 is Very Likely]

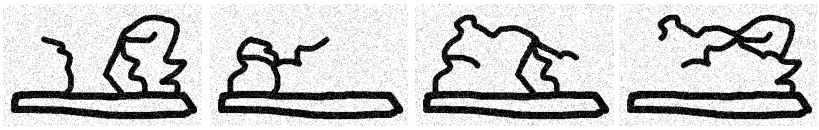
The participants were told that the questions were all independent, so that they would be free to rate an object as highly likely for both caves if they chose. Participants kept the training images during testing to avoid memory effects.

After they completed all questions they were debriefed and a photo was taken of the way they had arranged the study cards.

(a) *Seen*



(b) *Unseen*



(c) *Shuffled*

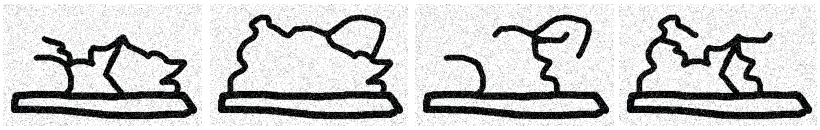


Figure 11: (a-c) Examples of types of test objects for Partition 1 A

Results

Figure 12 displays the results. Participants rated images from different test types significantly different (main effect of test type $F(3,36) = 97.28$, $p<0.001$ with sphericity corrections). The *seen same* ratings were significantly higher than the *seen other*, *unseen* and *shuffled* ratings ($t(15) = 15.58$, $p<0.001$; $t(15) = 9.00$, $p<0.001$; $t(15) = 11.32$, $p<0.001$, respectively). The *unseen* and *shuffled* scores were significantly different ($t(15) = 2.72$; $p<.02$). However, the *unseen* ratings were much closer to the *shuffled* than the *seen same* ratings. By analyzing pictures of how participants sorted the cards, we found two main arrangements: *integrated*, where the cards from both categories were lined up together (7 participants), and *separated*, where the cards were placed with significant space between them (5 participants). Also, four participants had no clear strategy (labeled *other*). For all three groups the difference between the *unseen* and *shuffled* images was not significant, but for the integrated group the difference between *unseen* and *shuffled* ratings was trending towards significant (*integrated*: $t(6) = 2.12$, $p = .08$; *separated*: $t(4) = 2.04$, $p = .11$; *other*: $t(3) = .94$, $p = .42$). Figure 12 shows the results from the three participant groups, along with representative examples of each card arrangement.

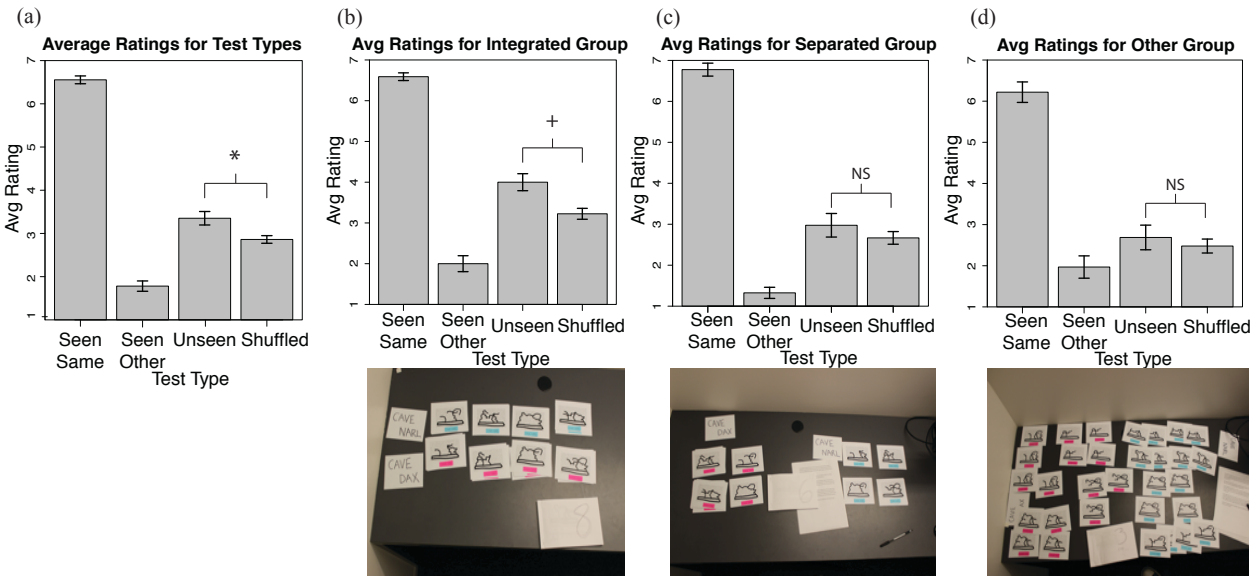


Figure 12: (a) Experiment 1 results. Results based on coding participants as (b) *integrated*, (c) *separated*, and (d) *other* groups. A representative example of each type of arrangement is under each graph of their results.

Discussion

Overall, the results support the label hypothesis and the IBP+. Whereas the IBF predicts that the *seen other* and *unseen* images should be rated as likely to belong to the category since they all share the same six parts, participants clearly rated both of those images as significantly lower than the *seen* images. There was a slight difference between the *unseen* ratings and the *shuffled* ratings, a result only predicted by the IBF. However, the IBF predicts a much higher rating of the *unseen* images, and the slight increase could have been caused by participants finding one or two independent features, as in fact the IBP+ did on many trials. The arrangements were analyzed in an attempt to see if strategy played a role in participants feature formation. We hypothesized that *integrated* participants might have been pushing the increase in *unseen* ratings: since they were looking at all images seemingly together, they should be more likely to find the six parts. Although we found no significant difference, the *integrated* participants were trending towards significance, and it is possible that we were only unable to find a significant difference due to the small sample size. If this is the case, then the *integrated* group would be acting like a weaker version of the hypothesis. Perhaps they did not generalize more fully because they only located some of the independent features or because they had learned the amounts of variation acceptable in the category in addition to the features. Only a larger sample size in the future could help answer these particular questions about card arrangement.

Although these results support the *label* hypothesis, there is an alternative explanation that does not involve feature formation: participants gave large ratings to observed objects and small ratings to unobserved objects (and those from the other category). Experiment 2 rules this possibility out.

Experiment 2: Control

In Experiment 2, both a Unitized and a Factorial set from a single partition were used during training. If participants are looking fro features, then they should find the six underlying parts and rate the unseen images higher for the Factorial cave. The results from the Unitized cave question provide a partial replication of Experiment 1.

Methods

Participants:

Seventeen Brown University students took part in this experiment either for course credit or for pay at a rate of $10 an hour. One participant's data was removed due to failing to complete the task.

Stimuli:

The stimuli for this experiment came from the same master object as Experiment 1. The Unitized sets used for training each partition were also the same as Experiment 1 (Unitized sets A and B). Each partition also had two Factorial sets, A and B. Factorial set A contained sixteen objects, including those from Unitized Set A and Unitized Set C—i.e. it was the set of all objects from the list of possible 20 objects excluding those from Unitized Set B (see Figure 13). It was always paired with Unitized Set A during study, and *unseen* questions for that test were the Unitized Set B questions. Factorial Set B contained sixteen objects including those from Unitized Sets B and C. For Factorial Set B, *unseen* test questions were images from Unitized Set A. Because all of the images from the Unitized cave also occurred in the Factorial cave, there were no true seen other images for the Factorial cave. Instead, all images rated in the *seen other* category for the Factorial cave are actually "seen both" images, seen in the Factorial cave but also in the Unitized cave. For the Unitized cave the *seen other* questions were images seen in the Factorial cave but not the Unitized cave. Training images were again printed on cardstock with attached labels.

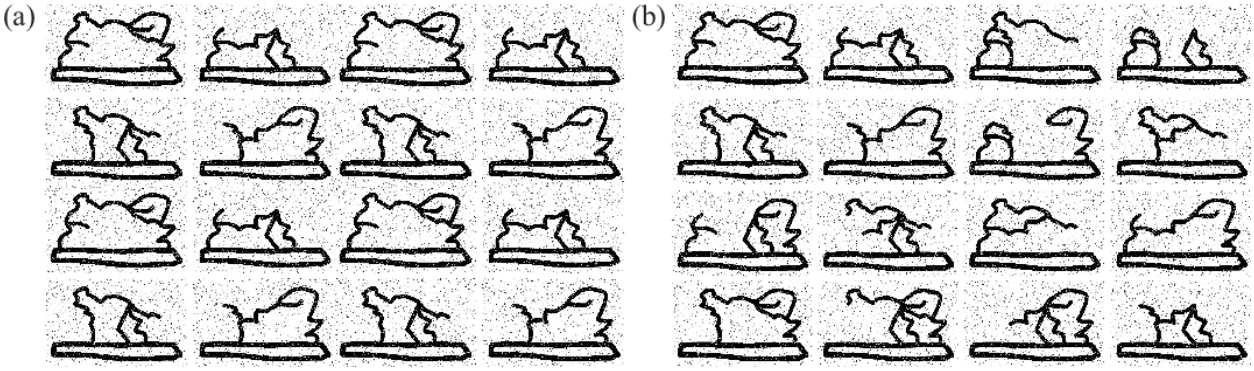


Figure 13: (a) Unitized set A and (b) Factorial set A

Design and Procedure:

This experiment had four independent variables. Because each unitized set was paired with a factorial set, each participant either had Partition 1 or 2, Sets A or B, order Factorial-Unitized or Unitized-Factorial, and test order 1 or 2. These four binary independent variables led to sixteen different counterbalanced conditions with one participant in each condition. The procedure was the same as Experiment 1.

Results

Figure 14 plots the results. Participants rated images from different test types significantly different (main effect of test type $F(3,24) = 50.13$, $p < .001$ with sphericity corrections, and this interacted with whether they were from the Unitized or Factorial cave ($F(3,24) = 19.53$, $p < .001$ with sphericity corrections). Participants rated the *seen* images from the Unitized cave significantly more likely to be in the Unitized cave than *seen other*, *unseen* and *shuffled* images ($t(15) = 8.47$; $t(15) = 10.71$; $t(15) = 12.47$, all $p < .001$). But, their Unitized cave ratings did not differ for the *unseen* or *shuffled* images ($t(15) = 1.64$, $p = .12$). Participants rated the likelihood of observing the *seen*, *unseen*, and *shuffled* images in the Factorial cave significantly different (*seen* and *unseen*: $t(15) = 3.91$, $p < .01$; *seen* and *shuffled*: $t(15) = 5.19$, $p < .001$; *unseen* and *shuffled*: $t(15) = 3.45$, $p < .01$).

The *unseen* and *shuffled* ratings in the Factorial cave were significantly larger than in the Unitized cave ($t(15) = 6.45$, $p < .001$; $t(15) = 6.15$, $p < .001$, respectively).

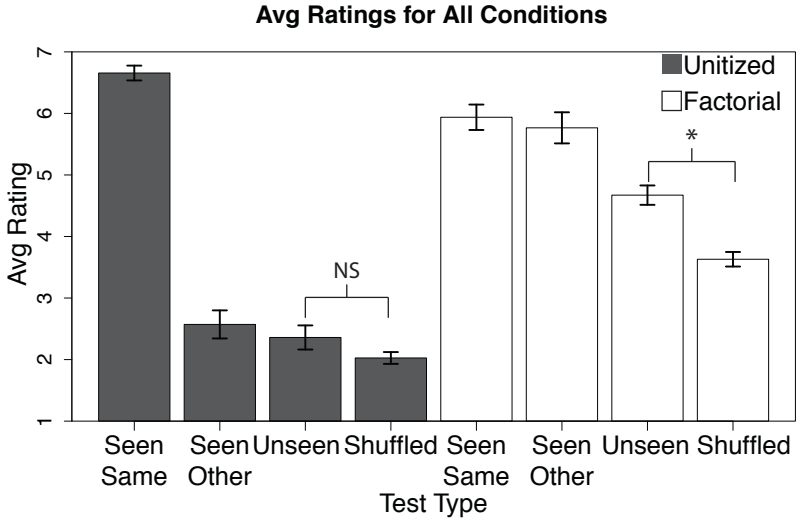


Figure 14: Results from Experiment 2

Discussion

These results support the *label* hypothesis. Participants rated the *unseen* images higher than *shuffled* images for the Factorial condition, indicating that they found the six parts. However, they did not extend these parts to the Unitized cave. Instead, they used the four given images as features for the Unitized cave and did not generalize to *unseen* images. In addition, the results suggest participants may also learn acceptable levels of variation: Participants gave larger ratings to the *unseen* and *shuffled* images in the Factorial cave than the Unitized cave.

The results are an imperfect replication of the data from Austerweil and Griffiths (2011). Their results showed either no significant difference or a very small difference between *seen* and *unseen* images for the Factorial condition (Austerweil & Griffiths, 2011). However, that result could have been caused by the poor Gestalt goodness of the shuffled images increasing the likelihood of better formed images. Although participants in this new experiment rated the *seen* and *unseen* images significantly differently for the Factorial cave, the also rated the *unseen* images to be significantly more likely than the *shuffled* images, but

not for the Unitized cave. This result, suggests that participants found the six underlying parts for the Factorial cave.

General Discussion

The results from participant data in this experiment are more consistent with the IBP+ model (and the *label* hypothesis) than the IBF model. The control data is especially telling. Even though participants recognized the six features for the Factorial category, they did not use these features for the Unitized category. Clearly, features are not always shared across categories. However, this is not necessarily a condemnation of the IBF or the *structure* hypothesis. The IBF model was created specifically for cases where participants were able to recognize shared features and the IBP+ model was not. Further, because of the way category labels are appended as extra information, the IBP+ places an artificial upper bound on the number of categories, and it is necessary for the IBP+ to know the number of categories *a priori*. If a new category is found, the objects must be relabeled and the process must start from the beginning. This reduces the psychological plausibility of the IBP+.

The result is especially surprising in light of the categorization literature, which assumes that category labels are treated as independent features only as children and as true "category markers" as adults (see Deng & Sloutsky, 2012). One possibility is that feature formation is inherently different than categorization, and therefore the processes are different. Given the intimate connection between categorization and feature formation, we would want more empirical support for this possibility. We discuss other possibilities below.

There may be a required closeness or shared relation between categories in order to share features. In Experiment 1, no instruction was given to find differences or similarities between the categories, and participants formed at least two clear strategies for organization: integrated and separated. Participants in the *integrated* group who looked like they may be sharing features or have found some independent parts may have felt compelled to find similarities, while others may have been searching for differences. If the relatedness of the categories is emphasized, such as if the two given categories belonged to an overarching category, then participants may be more likely to share features. This would actually be more similar to the example case of the categories TABLE and CHAIR, which share the overarching category FURNITURE. Arranging cards may serve as a catalyst for activating comparison mechanisms, which could be used in future experiments to understand the effect of comparison on feature formation. If arrangements have some effect on the sets participants are analyzing, then starting participants with pre-set arrangements similar to the integrated and separated arrangements should affect the patterns of generalization.

Even if participants were to share features across categories, there are some obvious problems with the psychological validity of the predicted IBF results. Given the way IBF currently performs, if the two groups share features, the model will not be able to distinguish between the groups when generalizing to new members—all objects with the appropriate features would belong in both categories. This includes the objects seen in the other set during training. In Experiment 1, participants very clearly rejected objects that had been seen in the other category. In Experiment 2, participants clearly rejected objects seen in the other category for the Unitized cave.

In addition, the ratings for *unseen* images were lower for the Unitized cave than the Factorial cave. This indicates that people may be using additional rules to organize their categories. Neither model currently accounts for the presence of possible additional hierarchical rules. The IBP model had a similar issue when originally proposed. Given a set of objects, the model was able to learn from covariation whether or not the two vertical lines were one feature (lines always move together) or two (lines can move independently). However, in the case of independently moving lines it was also then more confident than human participants that an object with one or three bars also belonged to the category. It appeared that human participants had also learned another rule that the IBP lacked about the number of features allowed in the category (Austerweil & Griffiths, 2011). A similar problem may be at hand for the current experiment. Even if participants had recognized the six features for both categories in Experiment 2, it is still apparent that the variation in the two categories is not the same. This leads to the conclusion that the unitized category may be restricted in the amount of variation acceptable. It is possible that by incorporating these rules into the IBF we will have a more accurate model and the *structure* hypothesis could be reconsidered for this case.

More experiments are necessary in order to make more definitive conclusions about the validity of either the IBP+ or the IBF. Ideally, more participants would be tested to draw conclusions about the potential arrangement strategies, and further experiments on the effect of shared categories and hierarchical rules would shed some light on this complex problem. The results from this experiment do show that features are not shared across categories in all cases, and that more work is needed to define the situations where features are shared and not.

References

Aldous, D. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII*, pp. 1–198. Berlin: Springer.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409–429.

Austerweil, J. L., & Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive Psychology, 63*, 173–209.

Austerweil, J.L., & Griffiths, T.L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review, 120*(4), 817-851.

Deng, W, & Sloutsky, V.M. (2012). Carrot eaters or moving heads: Inductive inference is better supported by salient features than by category labels. *Psychological Science, 23*(2), 178-186.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 1, 209-230.

Goodman, N. (1972). *Problems and projects*. New York, NY: Bobbs-Merrill.

Goldstone, R. L. (1998). Perceptual Learning. *Annual Review of Psychology, 49*, 585-612.

Griffiths, T. L., & Ghahramani, Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research, 12*, 1185–1224.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford, England: Oxford University Press.

Hoffman, D. D., & Richards, W. A. (1984). Parts in recognition. *Cognition, 18*, 65–96.

Murphy GL. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*(3), 289-316.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. (1998). Development of features in object concepts. Behavioral and Brain Sciences, 21, 1–17.

Schyns, P.G., & Murphy, G.L. (1994). The ontogeny of part representation in object concepts. *The Psychology of Learning and Motivation, 31*, 305-349.

Schyns, P.G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 23*(3), 681-696.