Categories and Feature Inference: Category Membership and a Reasoning Bias

Takashi Yamauchi (tya@psyc.tamu.edu)
Na-Yung Yu (dbskdud40@tamu.edu)
Department of Psychology, Mail Stop 4235
Texas A&M University, College Station, TX 77453

Abstract

This study investigated how the information about category membership influences the prediction of properties of category members. Results revealed that participants' response patterns became homogeneous and polarized when arbitrary noun labels carried category membership information. We suggest that this tendency arises because category membership information is represented like an abstract rule and triggers a reasoning bias.

Keywords: Categorical reasoning

Categories and inferences are two of the most common forms of organizing and generating new knowledge (Michalski, 1989). We create new categories to make predictions and analyses, and obtain new inferential knowledge on the basis of the conceptual categories we form. Categories such as medical diagnosis and biological taxonomies underscore the significance of categorization and inductive inference (Murphy, 2002).

How do we use categories for predictive inferences? Although many studies have documented the inductive potential of categorization, exactly *how categories modulate our inferential behavior remains unclear*. One theory suggests that noun labels, especially those related to natural objects such as *dog*, *cat*, and *tree*, create special expectations in an observer, and guide the person to make predictions in a way consistent with his/her expectations (Gelman, 2003). Another theory suggests that category information is no different from other regular attributes (Anderson, 1990; Osherson, et al., 1990; Sloman, 1993, 1998; Sloutsky, 2003). The information about category membership may draw more attention; yet, people interpret category membership as they do for other perceptual and conceptual attributes.

In this paper, we propose that category membership plays a special role in feature inferences, and molds people's inferential behavior in a way other regular attributes cannot. Specifically, we propose that the awareness of category membership generally creates a reasoning strategy and biases people's inferential behavior.

Consider a simple prediction task in which one infers the value of an unknown feature on the basis of another stimulus (Figure 1) (Murphy & Ross, 1994; Yamauchi & Markman, 2000). In one case, two stimuli have the same arbitrary label "monek" (Figure 1a); in the other case, two stimuli have different labels "moneke" and "plaple." In this circumstance, we think that people generally apply the following reasoning rules proportional to the extent to which the two labels carry the information about category membership: *Rule 1* – if two items belong to the same category, then the two items have characteristics in common; *Rule 2* – if two items belong to different categories, then the two items have different characteristics.

Undoubtedly, this reasoning strategy is erroneous because the members of natural categories are organized probabilistically, and the shared label does not necessarily guarantee shared features (and vice versa) (Rosch & Mervis, 1975). Moreover, psychological responses that we can observe in empirical studies are characteristically probabilistic. Therefore, such extreme "category-based" responses would rarely happen. However, we think that there is a cognitive bias to apply this "reasoning habit" in feature inferences when category information is transparent.

This reasoning strategy reflects the mutual-exclusivity constraint suggested by E. Markman (1989) and the psychological essentialism assumption suggested by Medin and Otorny (1989), and Gelman (2003). For example, the reason why shared labels lead to shared features is because a category is bound by some unknown or unknowable *essential* features, and these essential features generate other features. Likewise, two categories are viewed as mutually exclusive because they are bound by two sets of essentially different features.

We hypothesize that the mere presence of category labels promotes this rule-based reasoning strategy and generates *polarity* and *uniformity* in feature inference (e.g., Goldstone, 1994; Tajfel, 1963). For example, by applying this induction strategy, people accentuate the difference between two groups (i.e., polarity hypothesis), and discount perceptual variability of individual stimuli (i.e., uniformity hypothesis). We tested whether or not such reasoning biases would appear when the arbitrary labels carry category membership information.

Experiment

In our experiment, participants received pairs of a sample stimulus and a test stimulus one pair at a time (Figure 1), and predicted the feature value of a test stimulus on the basis of the sample stimulus. The stimuli were schematic illustrations of cartoon bugs, which were composed of 5 feature dimensions with binary values (Table 1).

Twenty test stimuli were presented twice. In one case, a test stimulus was paired with a sample stimulus that had the same label, and in the other case, the same test stimulus was paired with a sample stimulus that had a different label (Figures 1a & 1b).

Figure 1: A stimulus frame ((a) matched and (b) mismatched stimuli)

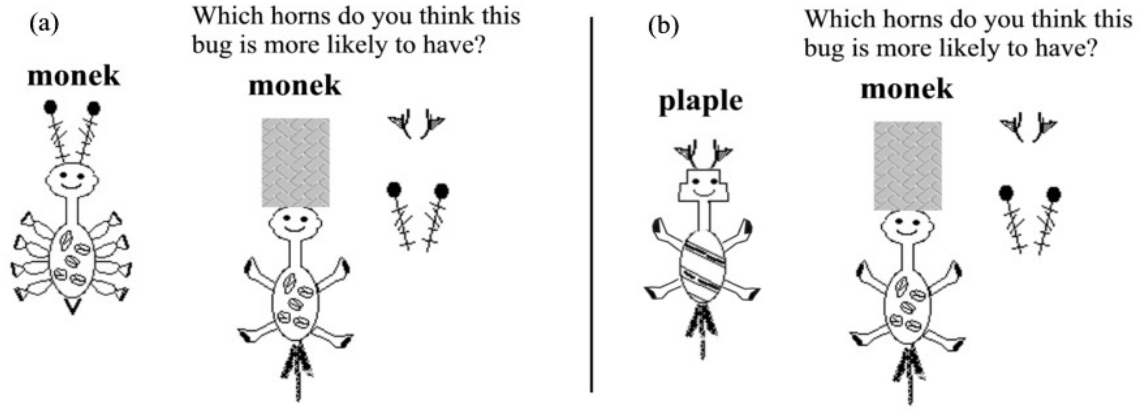


Figure 2: 2 sets of prototypes

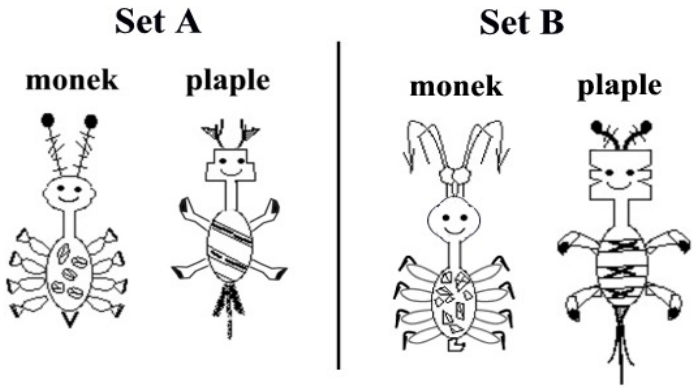


Table 1: Category Structure
(? indicates the features used for questions)

	horns	head	body	legs	tail	labels
monek	1	1	1	1	1	1
M1	?	1	1	0	0	1
M2	1	1	0	0	?	1
M3	1	0	0	?	1	1
M4	0	0	?	1	1	1
M5	0	?	1	1	0	1
plaple	0	0	0	0	0	0
P1	?	0	0	1	1	0
P2	0	0	1	1	?	0
P3	0	1	1	?	0	0
P4	1	1	?	0	0	0
P5	1	?	0	0	0	1

We manipulated the characteristic of "labels" solely in the instructions. The instructions in one condition characterized the two labels with respect to some arbitrary category membership information – a category condition; the instructions in the other three conditions characterized the same labels with respect to other arbitrary attribute information. We hypothesized that the extent to which the two labels convey category information polarity and uniformity biases would appear. That is, people exhibit a strong tendency to consider that two stimuli have the same features when they have the same labels, and two stimuli have different features when they have different labels (polarity hypothesis). Furthermore, the tendency to apply this rule uniformly over a variety of different stimuli increases as the two labels carry category membership information.

Participants & Materials A total of 112 undergraduate students were randomly assigned to one of four conditions: a category condition (*N*=30), a disease-attribute condition (*N*=27), a food-attribute condition (*N*=29), and an island-attribute condition (*N*=26).

Each stimulus was composed of 5 dimensions of binary features: (horns=long/short, head=round/angular, body=dotted/striped, legs=eight legs/four legs, tail=short/long) and a label (monek/plaple). Every test stimulus had 2 out of the 5 features consistent with the prototype of one category and 2 features consistent with the prototype of the other category (Table 1), and 1 feature was masked for an inference question. Ten test stimuli were created from the 2 prototypes of Set A, and the other 10 test stimuli were created from the 2 prototypes of Set B (see Figure 2). These 20 test stimuli were shown twice. In one case, a sample stimulus and a test stimulus had the same label (i.e., *match* condition – Figure 1a). In the other case, a sample stimulus and a test stimulus had different labels (i.e., *mismatch* condition – Figure 1b). In one version of stimuli, the prototypes of Set A were shown as sample stimuli (Figure 2); in the other version of stimuli, the prototypes of Set B were shown as sample stimuli. These prototypes were related to each other in their abstract appearance but the exact appearance of the two sets was different. For example, "monek" prototypes in Set A and Set B both have long horns,

a round face, a dotted body, eight legs and a short tail, but specific appearance of individual features were different. These two sets of stimuli were used to test the effect of perceptual variability in feature inference.

Procedure & Design Participants were shown a pair of a sample stimulus and a test stimulus on a computer screen, and were asked to select one of two feature values for the body part in question. One was consistent with the feature shown in the sample stimulus, and the other was inconsistent with the feature shown in the sample stimulus. Participants were instructed to make their decisions on the basis of the sample stimulus. Each participant received a total of 40 trials.

The design of the experiment was 4 (labeling characteristic – category, disease-attribute, food-attribute, island-attribute) × 2 (matching status – match vs. mismatch × 2 (feature set – same vs. different) factorial. *Labeling characteristic* was a between-subjects factor and this manipulation was made solely in the instructions that participants received. In the category condition, the two labels were characterized as representing two "types" of bugs. In the disease-attribute condition, the two labels were characterized as representing two kinds of "disease" that the bugs carry. In the food-attribute condition, the same labels were characterized as representing two kinds of "food" that the bugs eat regularly. In the island-attribute condition, the two labels were characterized as representing two different "islands" where these bugs live. All the other aspects of the experiment were identical across the four conditions. *Matching status* represents the matched/mismatched status of two labels displayed in a sample stimulus and in a test stimulus (Figures 1a & 1b). *Feature set* stands for the correspondence of the feature sets used to depict sample stimuli and test stimuli.

Table 2

	Match	Mismatch	Polarity
(same feature set)			
category	.81	.22	.59
disease	.71	.34	.37
food	.65	.36	.30
island	.60	.40	.20
(different feature set)			
category	.69	.32	.37
disease	.65	.38	.27
food	.62	.37	.25
island	.58	.37	.21

Note. The numbers in the cells represent the proportions of selecting the feature value consistent with the sample stimulus (i.e., "correct" responses).

Results: Polarity Effect The responses that were consistent with the sample stimulus were coded as "correct" responses. For example in Figure 1a, selecting the long horns was defined as "correct," and in Figure 1b, selecting the short horns was defined as "correct." To examine the impact of "polarity," we calculated "polarity scores" for each participant by subtracting the proportion of correct responses for mismatched stimuli from that for matched stimuli.

Overall, labeling characteristic and feature set did not affect participants' overall performance. There were no main effects of these two factors; *F*s<1.0. However, there was a significant interaction effect between labeling characteristic and feature set – $F(1, 108)=11.63$, $MSE=0.02$, $p<.01$. To identify the location of the interaction effect, we applied one way ANOVA separately to the two levels of feature set. This analysis showed that the performance in labeling characteristic differed primarily in the same feature set but not in the different feature set (Table 2). Given the same feature set, the mean polarity scores obtained from the four labeling conditions were significantly different; $F(3, 108)=4.75$, $MSE=0.168$, $p=.004$. Such a disparity was not observed in the different feature set; $F(3, 108)=1.03$, $MSE=0.128$, $p=.380$. To isolate the sources of the main effect, planned t-tests were applied to the data taken from the same feature set alone. The difference between the category condition ($M=.590$, $SD=0.384$) and the food-attribute condition ($M=.297$, $SD=0.408$) as well as the difference between the category condition and the island-attribute condition ($M=.20$, $SD=0.357$) was significant; (category vs. food) $t(57)=2.85$, $p=.02$, $d=0.741$; (category vs. island) $t(54)=3.85$, $p=.001$, $d=1.03$ (Bonferroni). The difference between the category condition and the disease-attribute condition ($M=.374$, $SD=.460$) was not significant; $t(55)=1.93$, $p=.177$, $d=.511$. Note that the failure to reach a significant level in this t-test came from the fact that the alpha level was adjusted with the Bonferroni method. The effect size of the two sample means was large ($d=.511$). Overall, a comparison between the category condition and the other three attribute conditions ($M=.29$, $SD=0.416$) revealed a significant disparity as well; $t(110)=3.42$, $p<.001$, $d=0.732$. These results showed that the mean polarity score in the category condition was significantly greater than that in the other three attribute conditions combined, suggesting that characterizing two labels with category membership information indeed polarized significantly subjects responses.

Attention weight and the polarity disparities. Did the polarity differences observed in the labeling conditions stem from different attention weights associated with the four types of labeling? Sloutsky suggests that feature inferences in young children are grounded in a "perceptual and attentional mechanism" that detects multiple similarities between stimuli (p. 247, Sloutsky, 2003). If the same similarity-matching mechanism is instrumental in our adult subjects, then the observed polarity differences should be explained merely by different level of attention that the four types of labels generated, rather than a reasoning bias per se. Following Sloutsky's suggestion, we examined if our results can be explained by a "perceptual and attentional" mechanism.

Let us assume that the probability of selecting a feature value of i in test stimulus X given sample stimulus Y is monotonically related to the similarity between X and Y:

$$P(X_i \mid Y) = \Phi(Sim(X, Y)) \quad \text{-- (1)}$$

where Φ is a probability density function that translates a

similarity value between *X* and *Y* into a unique probability value. In the Sloutsky model, the similarity between two stimuli, *X* and *Y*, is measured by (2):

$$Sim(X,Y) = W_{Label}^{1-L} S_{Vis.attr}^{N-k} \quad \text{-- (2)}$$

where *N* represents the total number of visual attributes in the two stimuli, *k* denotes the number of matching attributes between *X* and *Y*, $S_{Vis.attr}$ is an attention weight parameter associated with perceptual attributes, W_{Label} is an attention weight parameter for verbal labels. L represents the matching/mismatching status of labels. When two stimuli have the same label, then L=1, when the two stimuli have different labels, then L=0. Sloutsky and Fisher (2004) demonstrated that this function can account for young children's feature inferences accurately.

In our setting, the probability of selecting feature value i in test stimulus *X* given sample *Y* when *X* and *Y* have mismatching labels can be expressed in (3)-(6).

$$P_{mismatch}^{(type)}(X_i \mid Y) = \Phi(W_{Label\_type} S_{Vis.attr}^{N-k}) \quad \text{-- (3)}$$

$$P_{mismatch}^{(disease)}(X_i \mid Y) = \Phi(W_{Label\_disease} S_{Vis.attr}^{N-k}) \quad \text{-- (4)}$$

$$P_{mismatch}^{(food)}(X_i \mid Y) = \Phi(W_{Label\_food} S_{Vis.attr}^{N-k}) \quad \text{-- (5)}$$

$$P_{mismatch}^{(island)}(X_i \mid Y) = \Phi(W_{Label\_island} S_{Vis.attr}^{N-k}) \quad \text{-- (6)}$$

Because all participants in the four conditions in our experiment received the same stimuli, parameter $S_{Vis.attr}^{N-k}$ can have the same value across the four labeling conditions. Thus, the different polarity levels observed between the category condition and the three attribute conditions should have arisen from different values of W_{Label}^{1-L}, provided that the similarity-matching function is primarily responsible for the observed results. That is,

$$W_{Label\_type} < W_{Label\_disease}, W_{Label\_food}, W_{Label\_island} \rightarrow$$
$$P_{match}^{(type)}(X_i \mid Y) - P_{mismatch}^{(type)}(X_i \mid Y) >$$
$$P_{match}^{(disease)}(X_i \mid Y) - P_{mismatch}^{(disease)}(X_i \mid Y),$$
$$P_{match}^{(food)}(X_i \mid Y) - P_{mismatch}^{(food)}(X_i \mid Y),$$
$$P_{match}^{(island)}(X_i \mid Y) - P_{mismatch}^{(island)}(X_i \mid Y) \quad \text{--- (7)}$$

(7) indicates that the polarity difference between the category condition and the other attribute conditions was caused merely by different values of attention parameter *W*. (7) further implies that if the performance for mismatched stimuli is equivalent, attention weight *W* is also equivalent (e.g., $P_{mismatch}^{(type)}(X_i \mid Y) = P_{mismatch}^{(food)}(X_i \mid Y) \rightarrow W_{Label\_type} = W_{Label\_food}$ and see Appendix for proof). Following this reasoning, we equated participants' performance for mismatched stimuli over the four labeling conditions and examined if the observed polarity disparities would disappear. For this analysis, we first selected participants whose average scores for the *mismatched-same-feature set* stimuli were 0.4 or less (there were only four possible scores – 0.4, 0.3, 0.2, and 0 – that satisfy this criterion because each participant received 10 *mismatched-same-feature set* stimuli). We then calculated average accuracy scores for the *matched-same-feature set* stimuli over individual participants within each of the stratified levels (i.e., 0.4, 0.3, 0.2, and 0) (Table 2). These stratified average scores obtained in the three attribute conditions were compared to the average stratified scores obtained in the category condition by a paired *t*-test. This analysis showed that even after equating the performance for the mismatched stimuli, the polarity difference between the category condition and the other attribute-based conditions remained robust; *same feature set*; $t(14)=2.24$, $p<.05$, $d=0.574$; *different feature set*; $t(14)=3.06$, $p<.01$, $d=0.790$.

Table 3: Stratified comparisons of the performance for the matched stimuli

(same feature set)

	0	0.1	0.2	0.3	0.4
category	.96	.75	.56	.38	0
disease	.94	.77	.50	.03	.05
food	.97	.65	.42	.18	.20
island	.80	.65	0	.32	-.05

(different feature set)

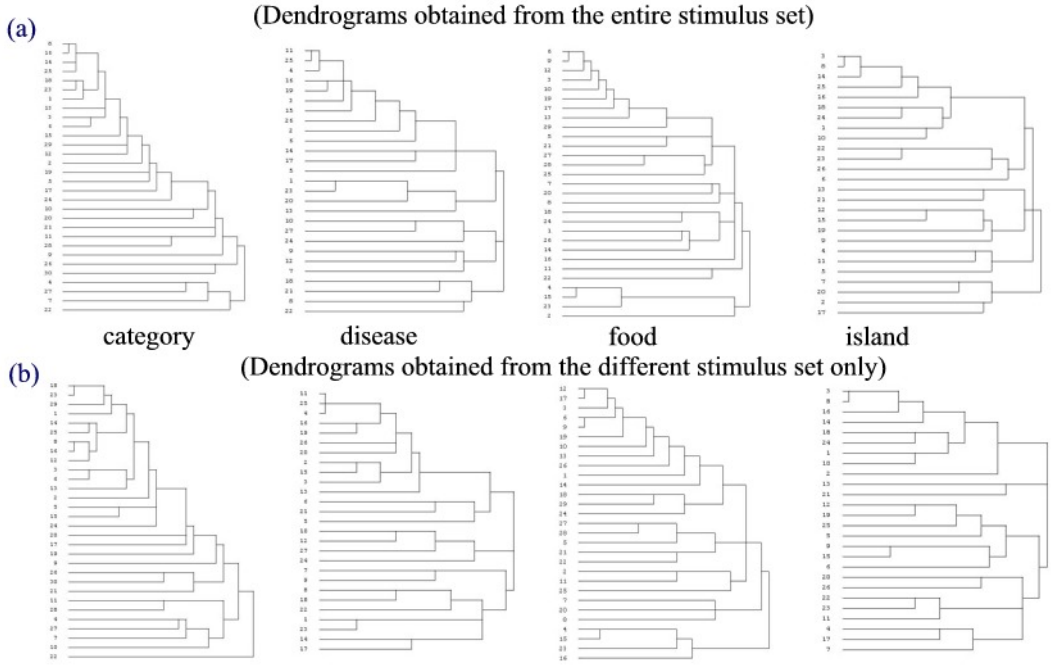
	0	0.1	0.2	0.3	0.4
category	.86	.70	.40	.28	.10
disease	.86	.57	.37	.03	-.05
food	.77	.65	.38	.22	.20
island	.80	.73	.30	.20	.05

Note. The numbers represent mean of polarity scores.

Clearly, it is unlikely that the disparity between the category condition and the other attribute conditions arose from the attention weight factor alone.

Uniformity Effect. The uniformity hypothesis suggests that when the two labels carry category membership information, the response patterns of individual participants become homogeneous. To test this hypothesis, we applied a cluster analysis and a correlation analysis. In our cluster analysis, we represented the entire responses of an individual participant with a vector of 40 dimensions (each dimension represents a response score (1or 0) obtained from one of 40 stimuli). We then applied a hierarchical cluster analysis by measuring squared Euclidian distances of individual vectors (see Figure 3 for dendrograms).

Figure 3: Cluster Analysis
(Dendrograms obtained from the entire stimulus set)



This analysis showed that an average dissimilarity distance of individual participants was 14.94 in the category condition, 17.65 in the disease-attribute condition, 18.37 in the food-attribute condition, and 18.88 in the island-attribute condition, suggesting that the response patterns obtained in the category condition were relatively homogeneous as compared to those obtained in the other attribute conditions. The four dendrograms (Figure 3) revealed that 80% of the participants (24/30) clustered in one group with a squared Euclidian distance of 18.21 in the category condition. When the same criterion 18.21 was applied to the other three conditions, only 48% of the participants (13/27) clustered in one group in the disease-attribute condition, 44.8% of the participants (13/29) clustered in the food-attribute condition, and 50% of the participants (13/26) clustered in the island-attribute condition.

Do these results reflect different levels of attention weights attached to verbal labels? Note that there were no statistical differences between the four labeling conditions in the *different feature set*. Thus, attention weight parameter *W* in the *different-feature set* stimuli should be roughly equivalent across the four labeling conditions (see (7) for this line of argument). In this regard, we applied the same cluster analysis solely to the responses obtained from the *different-feature set* stimuli. Even with this limited data set, the uniformity in the category condition was apparent (Figure 3b). Overall, 66.7% (20/30) of the participants in the category condition clustered in the same group with a squared Euclidian distance of 9.1 (the vectors in this analysis had 20 dimensions; therefore the Euclidian dissimilarity value is smaller). Given the same criterion 9.1, 51.9% (14/27) of the participants clustered in one group in the disease-attribute condition, 48.3% (14/29) of the participants clustered in the food-attribute condition, and 34.6% (9/26) of the participants clustered in one group in the island-attribute condition.

Correlation analysis. We examined the uniformity hypothesis with a correlation analysis as well. As in the cluster analysis, 40-dimensional vectors were constructed for individual participants. In each labeling condition, 26 vectors were selected randomly, and these 26 vectors were randomly divided into two groups of 13 vectors. The individual values of the 13 vectors were averaged over each dimension, yielding two group-vectors of 40 dimensions. We then measured Pearson's correlation coefficient between the two group-vectors. This procedure was repeated 1000 times in each labeling condition and the mean of correlation scores was calculated from a sample of 1000 (see Rosch & Mervis, 1975; for a similar analysis). This analysis shows that the response patterns observed in the category condition were highly correlated (*M*=.843, *SD*=.041), as compared to the other attribute conditions; (disease, M=.697, *SD*=.080), (food, *M*=.579, *SD*=.098), (island, *M*=.470, *SD*=.100). The same correlation analysis was applied only to the responses obtained from the *different feature set* stimuli. The overall results remained the same even for this limited data set; (category, *M*=.757, *SD*=.08; disease, *M*=.673, *SD*=.109; food, *M*=.570, *SD*=.122; island, *M*=.561, *SD*=.121).

Discussion The results from the experiment showed that inferential judgments that adult college students make were influenced significantly by the matched/mismatched status of labels when the labels convey category membership information. When a sample stimulus and a test stimulus had the same label, participants were more likely to predict that the two stimuli had other features in common. In contrast, when a sample stimulus and a test stimulus had different labels, participants tended to predict that the two stimuli had different features. This tendency was enhanced particularly when two arbitrary labels carried category membership

information as compared to when labels conveyed information about other attributes. We suggest that this bias arises because category membership information evokes a rule-like reasoning strategy. Although young children may use a similarity-based feature matching mechanism for feature inference (Sloutsky, 2003), adult subjects seem to employ a peculiar reasoning strategy specific to category membership.

Why do people employ a different reasoning strategy when category membership is transparent? Categories by default may be formed to subsidize inductive judgments, and for this reason, they may be ontologically distinct from other perceptual and conceptual attributes. For this reason, some mechanical rule-like feature predictions may be automatically triggered when stimuli convey information about category membership. It is also possible that the effect of category membership is context specific – it is learned later as one experiences how a wide variety of perceptual and conceptual groups work. Future studies have to examine the generality of the current finding as well as the source of this reasoning bias.

Appendix

Following the formulation by Sloutsky and Fisher (2004), we introduce (8) for the matched condition and (9) for the mismatched condition.

$$P(X_i \mid Y) = \frac{S^i}{S^i + S^j} \quad \text{--- (8)}$$

$$P(X_i \mid Y) = \frac{WS^i}{WS^i + WW'S^j} = \frac{S^i}{S^i + W'S^j} \quad \text{--(9)}$$

(8) shows the probability of selecting sample-consistent feature *i* (e.g., long horns in Fig. 1) in test stimulus *X* given sample *Y* when *X* and *Y* share a label. S^i represents the similarity between *X* and *Y* when *X*'s target feature has the value consistent with Y (long horns in Fig. 1a), and S^j is the similarity between X and *Y* when *X*'s target feature has the value inconsistent with *Y* (short horns in Fig. 1a). *W* is the attention weight for mismatching labels when S^i is given and *W*' is another attention weight for mismatching labels associated with S^j. Following the Sloutsky model, if $P^{(type)}_{mismatch}(X_i \mid Y) = P^{(island)}_{mismatch}(X_i \mid Y)$, then

$$\frac{S^i}{S^i + W'_{type} S^j} = \frac{S^i}{S^i + W'_{island} S^j}$$

$$W'_{type} = W'_{island}$$

References

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

Gelman, S. A. (2003). *The Essential Child: Origins of Essentialism in Everyday Thought.* New York: Oxford University Press.

Goldstone, R. L. (1994). Influence of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123*, 178-200.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction.* Cambridge, MA: MIT Press.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179-195). New York: Cambridge University Press.

Michalski, R. S. (1989). Two-tired concept meaning, inferential matching, and conceptual cohesiveness. In S. Vosniadou & N. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 122-145). Cambridge, MA: Cambridge University Press.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology, 27*, 148-193.

Murphy, G. L. (2002). *The big book of concepts.* Cambridge, MA: MIT Press.

Osherson, D. N., Smith, E. D., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category based induction. *Psychological Review, 97*, 185-200.

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

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology, 25*, 231-280.

Sloman, S. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology, 35*, 1-33.

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences, 7*, 246-558.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and Categorization in Young Children. *Journal of Experimental Psychology: General, 133*, 166-188.

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgment. *British Journal of Psychology, 54*, 101-114.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*, 776-795.