Title
Feature Inference: Tracking Mouse Movement

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 27(27)

ISSN
1069-7977

Authors
Kohn, Nicholas
Yamauchi, Takashi

Publication Date
2005

Peer reviewed

Feature Inference: Tracking Mouse Movement

Nicholas Kohn (nkohn@tamu.edu)
Takashi Yamauchi (tya@psyc.tamu.edu)
Department of Psychology, Mail Stop 4235
Texas A&M University, College Station, TX 77843 USA

Abstract

Past research suggests inductive judgments are made via simply assessing feature similarity (Osherson et al, 1990) while other research (Gelman & Markman, 1986) proposed that category labels convey information beyond other features. To further investigate these claims, we developed an online measure of decision-making. The present study examines how category labels affect inductive inferences by using a method akin to eye-tracking. The judgment results and the tracking data jointly support the view that category labels do affect inductive inferences in a way uniquely distinct from other feature information.

Conceptual categories such as *animal*, *vegetable*, and *furniture* are the basis of inductive inferences. For example, given a concept vegetable, we are able to infer its taste and appearance. Early in the 1990's, Osherson and his colleagues (Osherson et al., 1990, 1995) and Sloman (1993) made seminal observations that people make inductive judgments by assessing the similarity between items. On this view, the psychological strength of a conclusion is primarily determined by the similarity between a premise and a conclusion. For example, given a premise chimpanzees have disease X, the strength of a conclusion flamingos have disease X depends on the number of matching features between chimpanzees and flamingos. Following this important finding, a large number of studies have flourished in inductive judgments and impression formation research.

Although the work cited above provides convincing evidence that feature-based similarity is the main vehicle of human induction, the precise mechanism by which categories influence judgment processes is unknown. Is the strength of conclusions due to the number of matching features or is there another factor, something that will influence an inductive judgment above and beyond a similar feature? The present study proposes that a categorical label has properties that separate itself from other features. Specifically, a category label can influence one's decision process beyond the presence of similar features.

The idea of category labels having a special impact on inductive inference is not new. Gelman and Markman (1986) found that labeling dissimilar items with the same noun guides inductive judgments based on category membership rather than similarity between items. In this study, young children were presented with a triad of pictures. Each triad was constructed so that the third picture looked like one of the first two pictures, but was given the same label as the other picture. Children were told information about each of the two training pictures and were asked to infer which information applied to the third picture. Children based their judgments on category membership 68% of the time leading Gelman and Markman to conclude that category membership guides inductive judgments. While this finding supports their conclusion, it is not convincing. How exactly do labels guide inductive judgments? If participants were solely guided by the picture labels, then we would expect judgments based on category membership much more than 68% of the time. Furthermore, it is possible that the participants were simply matching labels on some trials so as to quickly solve the problem. Lastly, we do not know if these findings would hold true for adult participants. While it appears that there is an interaction occurring between feature similarity and category membership, we cannot be sure solely from the results of Gelman and Markman's study. It would advantageous to our investigation to know how participants make inductive judgments. For this purpose, we developed an online measure of how participants make inductive judgments when presented with stimuli that have experimentally manipulated feature similarity.

Many studies in feature inference have examined inductive judgments while manipulating the similarity of two stimuli (Gelman & Markman, 1986; Sloutsky & Fisher, 2004; Yamauchi, 2003). However, this paradigm focuses on participants' response patterns, and is not sufficient for completely investigating *how* people make inductive judgments. What we need is a real-time examination into how each participant makes an inductive judgment based on the information provided to him or her. To accomplish this, we used a new method that will be referred to as *mouse tracking*. In this experiment, stimuli were blurred to a point where visual recognition is impossible. In order to reveal the stimuli in a clearly visible form, participants had to move the mouse over the part of the stimulus they desired to see. Once they moved the mouse away from that area, it would be immediately blurred again. Thus, by using this experiment, we were able to trace a participant's viewing of stimuli and measure the reaction time (RT) spent on each of the features of a stimulus. This experiment is similar to eye-tracking experiments in its design and measures. Evidence has shown that the java-based computer program used for this study simulates eye movement behavior in a normal setting (Jansen et al., 2003). It should be noted that Payne, Bettman, and Johnson (1988) used a program called *Mouselab* to investigate strategy selection. However, *Mouselab* is not practical for online measures of inferences.

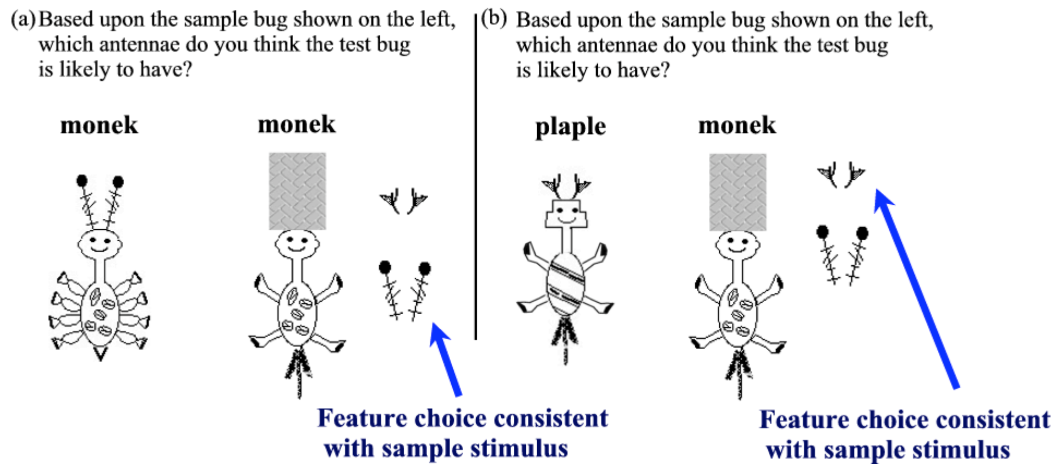

Figure 1: A stimulus frame ((a) matched and (b) mismatched trial)

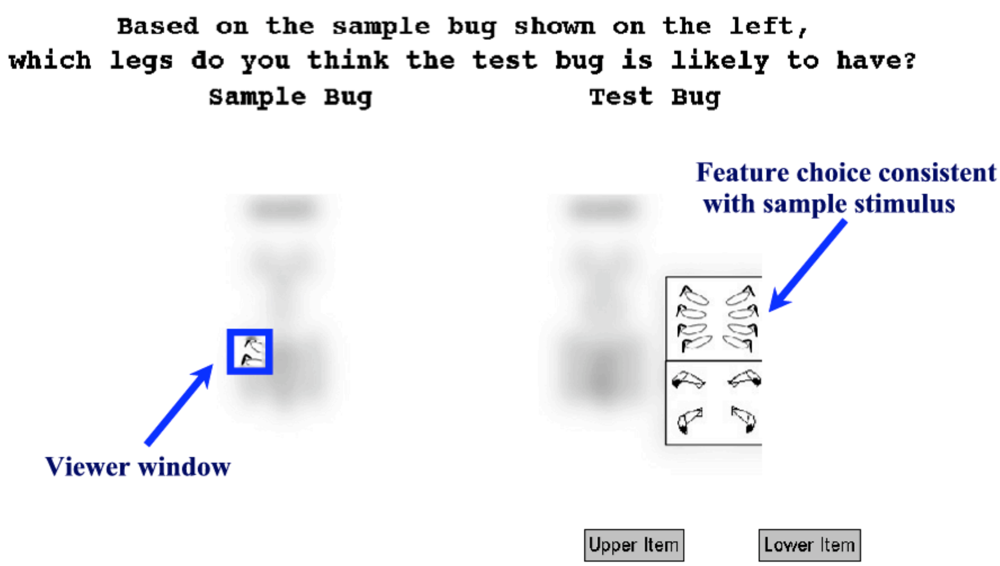

Figure 2: A sample of a trial in Exp. 2

Table 1: The category structure used in Experiments 1 & 2

(? Indicates the feature used for question)

	Antennae	Head	Torso	Leg	Tail	Label		Antennae	Head	Torso	Leg	Tail	Label
S1	**?**	1	1	0	0	1	S6	**?**	0	0	1	1	0
S2	1	1	0	0	**?**	1	S7	0	0	1	1	**?**	0
S3	1	0	0	**?**	1	1	S8	0	1	1	**?**	0	0
S4	0	0	**?**	1	1	1	S9	1	1	**?**	0	0	0
S5	0	**?**	1	1	0	1	S10	1	**?**	0	0	1	0
monek	1	1	1	1	1	1	*plaple*	0	0	0	0	0	0

In both of the present experiments, participants received pairs of stimuli: a sample stimulus and a test stimulus. Each of these was a fictional illustration of an insect. All test stimuli were produced from two sample stimuli by systematically replacing two of the five feature values. Thus, all test stimuli deviate from the corresponding sample stimulus in two feature dimensions (see Table 1).

The test stimulus had one feature missing and participants were asked to make a judgment about the missing feature. In addition to the two stimuli presented, two choices for the missing test feature were presented. Above each stimulus was a label. Depending upon the condition the participant was placed in, the instructions would state if this label referred to the category this stimulus belonged to or to the type of feature this stimulus had (e.g. type of wing). In Experiment 1, participants will view sets of these stimuli and corresponding labels. They will be asked to choose one of two feature choices for the test stimulus. Experiment 2 will repeat Experiment 1, but will be run using the mouse tracking program. The pairs of stimuli and their corresponding labels will be blurred, leaving only the inference question and the two feature choices clearly visible.

Previous studies indicated that people use category membership information like an abstract decision rule (Markman, 1989; Yamauchi, 2003). For example, if two items share a label, then people judge that the two items have characteristics in common; if two items have different labels, then people judge that the two items have different characteristics. This rule-based strategy states that participants' choice for the missing feature will be governed by the status of the labels. If the test stimulus has the same label as the sample stimulus, then the participant will select a feature for the test stimulus that is consistent with the sample stimulus. If the two labels are different, then the participant will select a feature for the test stimulus that is inconsistent with the sample stimulus. In Experiment 1, we will examine whether or not such extreme response patterns would emerge as a consequence of matched/mismatched status of labels. In Experiment 2, we predict that the extent to which labels guides participants' inductive judgments will be reflected in the reaction time spent viewing the stimuli.

Experiment 1

The goal of Experiment 1 was to investigate if we could find similar findings to Gelman and Markman's (1986) study. To accomplish this, we used the paradigm in which participants had to select a missing feature of a bug. Participants were shown this bug as well as a sample bug (all features present) to help guide their decision process. In addition to the bugs, labels were present above each of the two bugs. To control for any extraneous effects, the stimuli in the two conditions were identical. The only difference between the two conditions was the instructions given prior to the experiment. In the category condition, the arbitrary names "monek" and "plaple" were associated with two different *types* of the bugs, while in the feature condition, these two names were associated with two different shapes of wings.

Participants & Materials A total of 90 undergraduate students were randomly assigned to one of two conditions: a category condition (N=49) or a feature condition (N=41).

Twenty stimuli were devised for this experiment. These stimuli were schematic illustrations of cartoon bugs, which consisted of five dimensions of a binary feature (antennae = long or short, head = round or angular, torso = dotted or striped, legs = eight or four legs, tail = short or long) and a label ("Monek" or "Plaple"). The stimuli were created from two prototypes, which were defined to belong to one of two categories: monek or plaple. Table 1 summarizes the structure of the stimuli, and shows that each stimulus has three features consistent with the prototype of the corresponding category as well as two features consistent with the prototype of the other category.

Each trial contained a sample bug and a test bug. The test bugs had one of the five body features missing. In each trial, an inference question was presented in the following format: "Based on the Sample Bug shown on the left, which *FEATURE* do you think the Test bug is likely to have?" In the actual questions, *FEATURE* was replaced with one of the five feature terms – antennae, head, torso, legs, or tail. The target feature of the test bug was covered by a mask. Two depictions of a feature were presented as choices to answer the inference question. One of these feature depictions was consistent with the sample bug and the other is inconsistent with the sample bug. Participants indicated their choice of feature depiction by clicking the corresponding button.

Procedure & Design At the beginning of the experiment, each participant was given one of the two instruction sheets. This divided participants into the category condition or the feature condition. Category condition instructions indicated that labels shown above the bugs referred to the category the bug belonged to. Feature condition instructions indicated that the labels referred to the type of wing this bug has. The task of the participants was to answer 20 inference questions. For each trial, participants were shown a pair of sample and test stimuli on a computer screen, and were asked to select one of two feature values for the body part in question (Figure 1). Ten stimuli were shown twice – once paired with the corresponding sample stimulus (i.e., the sample and test stimuli had the same label), and once paired with the sample stimulus with the opposite category (i.e., the sample and test stimuli had different labels). The order of presenting trials was determined randomly for each participant. The dependent measure of this experiment was the proportion of inference question responses that select the feature value consistent with the sample stimuli. The design of the experiment was 2(instruction condition: category vs. feature) x 2(label status: matched vs. mismatched) factorial. Label status refers to the two types of trials: matched (stimuli labels match), and mismatched (stimuli labels are mismatched).

Results and Discussion As predicted, characterizing the two arbitrary names as category labels polarized participants' response patterns considerably in the category condition. In contrast, such extreme response patterns were absent when the same arbitrary names were associated with features in the feature condition (Table 2).

There was a significant interaction effect between instruction condition and label status; $F(1, 88)=20.16$, $MSE=0.06$, $p<.01$. Given the category condition, the disparity of performance between the matched and mismatched trials was 0.52, and given the feature condition, the disparity between the matched and mismatched trials was 0.24. This difference was significant; $t(88)=3.3$, $p<.01$, indicating that the matched label status influenced the performance in the category condition more so than that in the feature condition. There was a significant main effect of label status; $F(1,88)=70.75$, $MSE=0.06$, $p<.01$. The main effect of condition was not significant; $F(1, 88)= 2.36$, $MSE=0.03$, $p=.13$.

Table 2: Results from Exp. 1

	Category	Feature
Matched	.792 (.034)	.644 (.037)
Mismatched	.273 (.043)	.412 (.047)
Disparity Score	.518 (.060)	.232 (.067)

Note. These numbers are means and standard errors – enclosed in parentheses.

The results from Experiment 1 are consistent with the hypothesis that category labels, unlike other features, elicit an abstract rule-like decision strategy. There are at least two interpretations for these results. First, these results reflect participants being primarily guided by category information in their decision-making. Second, category labels might have influenced participants' response patterns just like other features because these verbal labels were simply more salient than other features. In this regard, it is difficult to infer the underlying *decision processes* simply from the results in Experiment 1. Experiment 2 addressed these problems.

Experiment 2

Experiment 2 was designed to replicate Experiment 1, but to also further investigate the mechanisms of inductive inferences. Experiment 1 provided evidence that participants were using a label-based strategy in their inference choices. However, we do not know the extent to which labels guided their decision process. If participants in the category condition solely relied on labels when choosing the missing feature, then we would have observed a larger difference between matched trials and mismatched trials. It is clear that more than just labels are affecting participants' induction process. Furthermore, it is also possible that participants in Experiment 1, as well as in other studies (Gelman & Markman, 1986; Sloutsky & Fisher, 2004), used the labels as a shortcut to complete the experiment quickly. Therefore, we need an online measure of how participants make inferences. Experiment 2 attempts to accomplish this by using an alternate form of eye-tracking, which we call *mouse tracking*. By measuring how long a participant spends viewing a feature, we can estimate how much that feature is guiding their decision process. Experiment 2 will use the same conditions and stimuli found in Experiment 1, but will present the trials using our mouse tracking program. We expected to find the same response patterns in the category condition and feature condition found in Experiment 1. Participants in the category condition will spend substantially more time viewing labels than participants in the feature condition. In addition, if the labels are used merely as shortcut to make a response quickly, then the overall response time in participants in the category condition would be significantly shorter than those from the feature condition.

Participants & Materials A total of 71 undergraduate students were randomly assigned to one of two conditions: a category condition (N=36), and a feature condition (N=35).

The stimuli for this experiment were the same illustrations of fictional bugs used in Experiment 1. To simulate eye-tracking equipment, a java-based Restricted Viewer program was used (Jansen et al., 2003). This program uses blurred images on the screen that can only be visibly revealed by moving a small viewer window using the mouse (see Fig 2). Thus, to view the entire stimuli image, the user must move the mouse around the screen. This program tracks the movement of the window across the images as well as the amount of time spent at each location. The size of the viewer window for this experiment was 47 x 47 pixels and the motion blur rate was set at 100.

Our Restricted Viewer program presented two stimuli (sample bug and test bug) that were blurred beyond visible recognition. Also burred was the label (either *Monek* or *Plaple*) appearing above each of the bugs. To create the blurring effect, four different versions of each stimuli were created, each treated with the Gaussian blur filter in Adobe Photoshop (with a radius of 5, 8, 11, or 14 pixels). These four images, combined with an unblurred version, were used by the program to create the desired effect. Above the two blurred stimuli, an inference question was presented in the same format as in Experiment 1. Two depictions of a feature were presented as choices to answer the inference question. One of these feature depictions was consistent with the sample bug and the other is inconsistent with the sample bug. Participants indicated their choice of feature depiction by clicking the corresponding button.

As in Experiment 1, the category and feature conditions differed only in the instruction sheet that was given to participants.

Procedure & Design. At the beginning of the experiment, each participant was given one of the two instruction sheets. Next, participants were given a practice trial with instructions on how to use the mouse cursor in the Restricted Viewer program to reveal the blurred images in order to answer the trial's inference question. Following this tutorial, participants completed the twenty experimental trials at their own pace.

Experiment 2 was also designed with a 2 (instruction condition: category vs. feature) x 2 (matching status: matched vs. mismatched) factorial. For mouse tracking, RT data was analyzed using a 2 (stimulus type: sample vs. test) x 2 (label status: matched vs. mismatched) x 2 (mouse location: label vs. body part) x 2 (instruction condition: category vs. feature) factorial. Mouse location refers to where the participant was focusing the viewer window.

Dependent measures for this experiment were the amounts of time a participant spent viewing features of each bug and *consistency* (the percentage of occurrences a participants answered the inference question using a feature that is consistent with the sample bug). Using the RT and mouse location data collected from the Restricted Viewer program, we could analyze how much time participants spent viewing the individual features and label of the sample and test stimuli.

Results & Discussion In this experiment, we eliminated the data from participants whose total time spent viewing the bug body parts and labels was less than 800ms. While this can be viewed as an arbitrary number, we reasoned it was impossible to answer each question adequately in less than 800ms. This left us with 23 participants in the category condition and 22 in the feature condition. To make sure that the response patterns obtained in the restricted viewer program were analogous to those obtained in Experiment 1, we first analyzed the proportion of selecting the feature consistent with the sample stimulus, a 2 (label status) x 2 (instruction condition) ANOVA was performed (Table 3). This revealed that the main effect of label status was significant [$F(1,43)=15.15$, $p<.01$, $MSE=0.05$], as was the interaction between label status and instruction condition [$F(1,43)=4.23$, $p=.04$, $MSE=0.05$]. To examine if a rule-like decision was being used in choosing a feature, a t-test examined the disparity between matched and mismatched trials between the two conditions. The category condition ($M=0.265$, $SD=0.352$) was significantly higher than the feature condition ($M=0.082$, $SD=0.230$); $t(43)=2.057$, $p<.05$. Although the impact of label status was ameliorated considerably in this experiment, this significant difference appears to support the theory that the category condition has invoked a rule-like strategy in participants' decision-making.

Table 3: Results from Exp. 2 (Means and Standard errors)

	Category	Feature
Matched	.583 (.052)	.514 (.039)
Mismatched	.317 (.046)	.432 (.042)
Disparity Score	.265 (.073)	.082 (.049)

To investigate how participants made their decisions, we now turn to the movement analysis. In this analysis, we predicted that there would be more time spent viewing labels in the category condition. The average time the participants spent viewing the labels and bodies of the stimuli can be found in Table 4. A 2 (stimulus type: sample vs. test) x 2 (label status: matched vs. mismatched) x 2 (mouse location: label vs. body part) x 2 (instruction condition: category vs. feature) ANOVA was performed on average viewing time. The mouse-location versus instruction-condition interaction was significant [$F(1,43)=4.40$, $p=.04$, $MSE=2913562.13$] as was the mouse location versus stimulus type interaction; $F(1,43)=31.75$, $p<.01$, $MSE=493653.67$. Participants in the feature condition viewed the body parts ($M=12614.59$, $SD=4567.006$) longer than the labels ($M=1016.45$, $SD=1089.474$); $t(40)=11.59$, $p<.00$. Body parts ($M=10225.91$, $SD=4580.228$) were also viewed longer than labels ($M=1647.35$, $SD=1120.805$) in the category condition; $t(44)=8.72$, $p<.00$. This was to be expected as body parts occupy much more stimulus area than the labels. Two main effects were also found from this ANOVA. Participants viewed the sample stimulus longer than the test stimulus [$F(1,43)=49.23$, $p<.01$, $MSE=556049.76$] and spent more time viewing the body parts of the stimuli than the labels; $F(1,43)=196.39$, $p<.01$, $MSE=2913562.13$. We can assume that the greater viewing RT of the sample stimulus reflects participants using the test stimulus only as a loose background of knowledge in which to do inductive inferences on the sample stimulus. Interestingly, the two conditions did not differ in terms of the overall times that participants used to make responses ($t(43)=0.15$, $p=.88$), suggesting that labels in the category condition did not work as a simple shortcut to complete the experiment. Rather, the two labels were used to assess other underlying features and to make inference judgments.

Lastly, to judge the extent labels were used by participants in making their feature decisions, we computed a statistic of the percentage of the entire stimulus RT spent viewing the label (label / (label + body parts)). A 2 (stimulus bug) x 2 (label status) x 2 (instruction condition) ANOVA on this percentage revealed that only the main effect of condition was significant; $F(1,43)=5.22$, $p=.03$, $MSE=0.07$. As expected, participants in the category condition (16%) spent a greater percentage of time viewing the label than participants in the feature condition (8%) (see Fig. 2).

Table 4: Mean RT Viewing Stimulus Feature (ms)

	Category	Feature
Sample Bug		
Matched Trials		
Label	441.48	316.95
Body Parts	3116.61	3712.68
Mismatched Trials		
Label	517.91	324.00
Body Parts	3076.48	3452.68
Test Bug		
Matched Trials		
Label	305.26	151.14
Body Parts	2090.96	2776.82
Mismatched Trials		
Label	382.70	224.36
Body Parts	1941.87	2672.41

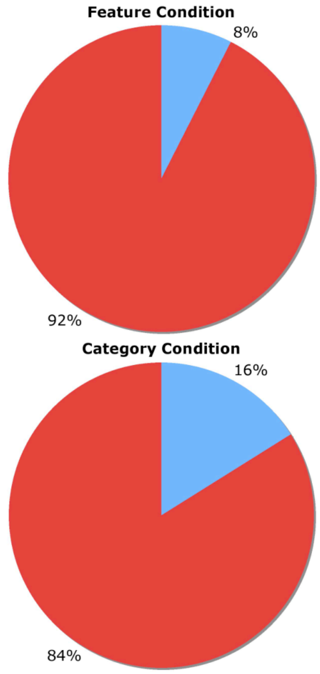

Figure 3: Percentage of viewing RT for labels and body parts

General Discussion

In the two experiments, we have investigated the role of category labels in feature inference. Specifically, we measured participants' mouse movement patterns and the way participants make feature inferences. There are a couple of important findings drawn from the results of the two experiments. First, the results from the first experiment showed that the matched/mismatched status of labels creates an extreme response pattern when two labels were characterized with category membership. Second, Experiment 2 showed that this extreme response pattern arose from the fact that participants in the category condition were using labels to a larger extent than participants in the feature condition.

Exactly, how did participants use labels to make judgments? First, it appears that the information about category membership provides more than an expedient means to make judgments. Category labels are not used merely to obtain a shortcut of responses, a quick and easy "decision-rule" per se. Rather, the category labels seem to redirect participants to assess other underlying features. Based on the mouse movement data it appears that an interaction is occurring between the labels and the other features of the stimuli. When labels refer to just feature information, participants do not appear to treat them any differently than other features of the stimuli. However, when the labels indicate category membership, participants have a different type of label–feature interaction. As illustrated in Figure 3, participants did not significantly spend more time viewing stimuli in one condition. Rather they allocated more time to viewing labels in the category condition. It appears, participants use the category membership to make different comparisons between the two stimuli and consequently make a different inductive inference. This supports prior research that category labels have characteristics separate from other features (Gelman & Markman, 1986; Yamauchi, 2003).

This experiment has not only shed light on how people use category membership and other features to make inductive inferences, but it has also demonstrated a new method for studying decision-making research. Granted, we cannot be totally sure what the participant is thinking when they view stimuli. However, the combination of decision responses and RT can jointly support conclusions drawn from the data. Using this mouse tracking software, we hope to perform future experiments investigating mechanisms of inductive inferences, category learning, and reasoning.

References

Gelman, S., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*, 183-209.

Jansen, A. R., Blackwell, A. F., & Marriot, K. (2003). A tool for tracking visual attention: The restricted focus viewer. *Behavior Research Methods, Instruments & Computers, 35,* 57-69.

Markman, E. M. (1989). Categorization and naming in children: Problems of induction. 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.

Osherson, D., Smith, E. E., Shafir, E., Gualtierotti, A., & Biolsi, K. (1995). A source of Bayesian priors. *Cognitive Science, 19*, 377-405.

Payne, J.W., Bettman, J.R., & Johnson, E.J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 534-552.

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

Sloutsky, V.M., & Fisher, A.V. (2004). When development and learning decrease memory: Evidence against category-based induction in children. *Psychological Science, 15*(8), 553-558.

Yamauchi, T. (2003). Dual processes in the acquisition of categorical concepts. In R. Altman & D. Kirsh (Eds.), The Proceedings of the 25th Annual Meeting of the Cognitive Science Society, Boston: Cognitive Science Society, 1259-1264.