

Tracking mouse movement in feature inference: Category labels are different from feature labels

TAKASHI YAMAUCHI, NICHOLAS KOHN, AND NA-YUNG YU
Texas A&M University, College Station, Texas

In this article, we examine the role of category labels in inductive inference. Some leading research has suggested that information about category membership works just like any other feature in categorical inductions, whereas other research has proposed that the influence of category membership on induction goes beyond that of other features. To investigate these claims further, we developed an online measure of judgments that is akin to eyetracking. The judgment results and the mouse-tracking data jointly support the view that category labels do affect inductive inferences in a way distinct from that for feature information. When arbitrary labels conveyed category membership information, participants viewed these labels more often and earlier in a trial, in comparison with cases in which the same labels conveyed non-membership information. Our results suggest that category membership information works like a guide for inference. An ecological rationale for this induction strategy is also discussed.

Consider the following sentence: *James is a Harvard-educated carpenter*. By assigning James this label of a *Harvard-educated carpenter*, we are able to elicit many inductive inferences about James that go beyond a man who is a carpenter who received an education from Harvard. Perhaps he is a man who values an education but prefers a simple lifestyle. Or perhaps James went to Harvard due to family pressure but is now rebelling against his father (Kunda, Miller, & Claire, 1990). In the present study, we attempted to investigate how category labels can influence inductive judgments.

Categorical noun labels can influence inference differently than does other feature information. However, contemporary theories of categorical induction make little distinction between category labels and feature labels. For example, in the model proposed by Sloman (1993), the strength of an argument is assumed to depend on (1) the extent to which a premise category and a conclusion category have attributes in common (similarity) and (2) the extent to which a premise category spans a conclusion category (coverage; hereafter, we will call these two factors *attribute-based similarity*). Indeed, major research in categorical inference has evolved around specifying the boundary conditions in which these two factors are applicable (Heit & Rubinstein, 1994; Medin, Coley, Storms, & Hayes, 2003; but see Ross, Gelman, & Rosengren, 2005). This approach is also prevalent in other fields, including judgment and decision making (e.g., representative heuristics; Kahneman & Tversky, 1973; Shafir, Smith, & Osherson, 1990; Tversky & Kahneman, 1983) and stereotyping and impression formation research (Allport, 1954; Duckitt, 1992; Hamilton & Sherman, 1994; Kashima,

Woolcock, & Kashima, 2000; Kunda & Thagard, 1996; Stangor, 2000).

Implicit in these studies is an assumption that category membership can be represented by a collection of features or individual instances (Anderson, 1990; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1986; Sloutsky, 2003; see also A. B. Markman & Ross, 2003, for a different approach) and that categorical noun labels simply point to the content of the category. In this view, category labels work just like other feature labels (Anderson, 1990).

In this study, we investigated the inductive potential of noun labels from a different perspective. We think that category labels have a special property that distinguishes them from other features and that they influence inductive inference differently from other attribute information. Unlike other features, category labels guide induction by leading an observer to actively search for similarities and differences among objects. In other words, rather than being a piece of information to be added as another feature, category labels initiate an inference process (Gelman & Coley, 1990; Gelman & Heyman, 1999; Yamauchi, 2005; Yamauchi & Markman, 2000). In the two experiments reported in this article, we investigated this hypothetical property associated with category labels. Specifically, we contrasted the process of predicting unknown features when arbitrary labels carried category membership information and when the same labels carried non-membership information.

In two experiments, participants received pairs of stimuli, a sample stimulus and a test stimulus, and were instructed to make a prediction about a test stimulus on

T. Yamauchi, tya@psyc.tamu.edu

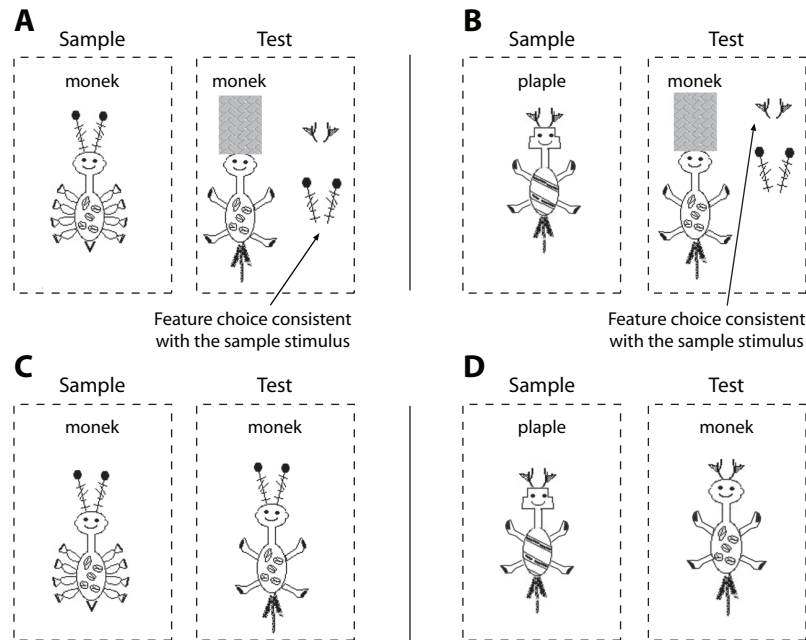


Figure 1. Examples of two stimulus frames for matched (A) and mismatched (B) trials of the inference task and for matched (C) and mismatched (D) trials of the similarity judgment task. In the matched trial, the sample and test stimuli had the same label; in the mismatched trial, the sample and test stimuli had different labels. The sample stimuli were the prototypes of the test stimuli (see Table 1).

the basis of a sample stimulus (Figures 1A and 1B; for a similar procedure, see Murphy & Ross, 1994; Yamauchi & Markman, 2000). These stimuli were fictional illustrations of an insect. All the test stimuli were produced from two sample stimuli by systematically replacing two of the five feature values (Table 1). The test stimulus had one feature missing, and the participants were asked to make a judgment about the missing feature. Two choices for the missing test feature were shown, and above each stimulus was a label, which conveyed either category membership information or feature information.

The main manipulation of this study was the instructions that the participants received. Depending on the condition they were assigned to, the instructions would state whether a label referred to the *category* a stimulus belonged to or to a *feature* a stimulus had (e.g., *the label represents the shape of wings*). We investigated how these different characterizations of verbal labels would influence the participants' response behavior. The manipulation of the presence of categorical versus feature information was done in this manner in order to ensure that the participants in both conditions viewed exactly the same stimuli.

Many studies in which inductive judgments have been examined have done so by manipulating the similarity between the stimuli and then analyzing the responses made by the participants (Gelman & Markman, 1986; Sloutsky & Fisher, 2004; Yamauchi & Markman, 2000). However, this paradigm is not sufficient for investigating *how* people make inductive judgments. In order to make a real-time examination into a judgment process, we used a new method that will be referred to as *mouse tracking*. In Experiment 2,

the stimuli were blurred to a point where visual recognition was impossible. In order to reveal the stimuli in a clearly visible form, the participants had to move the computer cursor/mouse over the part of the stimulus. Once they moved the mouse away from that area, it would be immediately blurred again (Figure 2). Thus, by using this technique, we were able to trace a participant's viewing of the stimuli and measure the time spent on each of the features of a stimulus. Our mouse movement measure is similar to eye-tracking procedures in its design and analyses. Evidence

Table 1
Category Structure Used in Experiments 1 and 2

Stimuli	Features					Labels
	Horns	Head	Body	Legs	Tail	
Test 1a	?	1	1	0	0	1
Test 2a	1	1	0	0	?	1
Test 3a	1	0	0	?	1	1
Test 4a	0	0	?	1	1	1
Test 5a	0	?	1	1	0	1
Sample A	1	1	1	1	1	1
Test 1b	?	0	0	1	1	0
Test 2b	0	0	1	1	?	0
Test 3b	0	1	1	?	0	0
Test 4b	1	1	?	0	0	0
Test 5b	1	?	0	0	1	0
Sample B	0	0	0	0	0	0

Note—All test stimuli were produced from two sample stimuli, Sample A and Sample B, by switching two of the five feature values. “?” represents an inference question given in an inference trial. (1, 0) = horns (long, short), head (round, angular), body (dotted, striped), legs (eight, four), tail (short, long), and labels (monek, plaple).

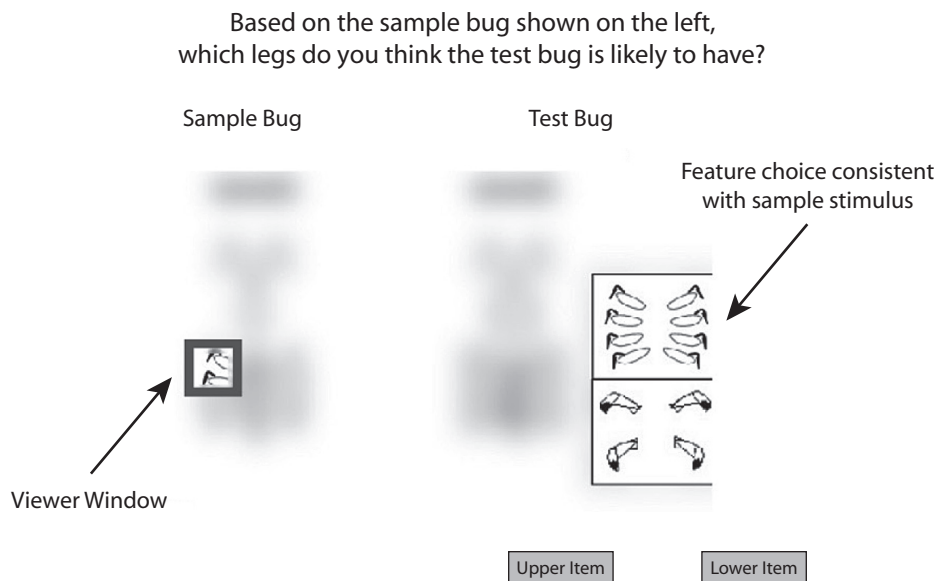


Figure 2. A sample of a stimulus frame of the restricted viewer program used in Experiment 2. The sample (left) and test (right) stimuli were blurred, and the viewer window moved as the participant moved the mouse. A small portion of the stimuli could be seen through the viewer window. In actual trials, the two prediction features shown next to the test stimulus were also blurred.

has shown that the Java-based computer program used for this study simulates eye movement behavior in a normal setting (Jansen, Blackwell, & Marriott, 2003).

Predictions

The predictions for the two experiments arise from a view that category labels play a guiding role in feature inference and helps integrate diverse features (E. M. Markman, 1989; Yamauchi, 2005; Yamauchi & Yu, 2005). The following two variables were assessed to test this idea. First, we measured the proportions for selection of the feature value consistent with the sample stimulus (consistency score) when the sample and the test stimuli had a common label (matched condition; Figure 1A) and when they had different labels (mismatched condition; Figure 1B). Our prediction was that the participants' consistency scores would rise when the two labels were matched and would decrease sharply when the two labels were mismatched (see Corneille, Klein, Lambert, & Judd, 2002; Tajfel & Wilkes, 1963). We expected that this tendency would increase sharply when the two arbitrary labels carried category information and would remain robust even after Sloman's (1993) similarity and coverage factors were controlled.

Second, in Experiment 2, we compared the frequency with which and the order in which verbal labels and other components were viewed. It was predicted that when two labels carried category membership information, category labels would be viewed more often, particularly at the beginning of a trial.

EXPERIMENT 1

Experiment 1 was an initial investigation into the role of categorical information in inductive inferences. We

examined whether category information would have an effect on inferential judgments above and beyond that of attribute-based similarity (similarity and coverage). To control the effect of similarity information, the participants carried out a similarity judgment task after a feature inference task. The participants' assessments of the similarity between the sample and the test stimuli were used to control the influence of perceived similarity between stimuli. To control the effect of coverage, two pilot studies were carried out (see the Appendix). To control other 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 labels *monek* and *plaple* were associated with two different *types* of bugs, whereas in the feature condition, these two names were associated with two different *shapes* of wings.

Method

Participants

A total of 133 undergraduate students participated in this experiment for course credit. They were randomly assigned to one of two conditions: a category condition or a feature condition. The data for 5 participants were removed because they responded exclusively with one of the two designated keys (binomial distribution test, $p < .001$). Altogether, the data from 128 participants (category condition, $n = 70$; feature condition, $n = 58$) were analyzed.

Materials

Ten stimuli were devised for this experiment. These stimuli were schematic illustrations of cartoon insects, which consisted of five dimensions of a binary feature (horns, long or short; head, round or angular; body, 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 belonged to one of two categories: *monek* or *plaple*. Each test stimulus had two features consistent with the

prototype of the corresponding category and two features consistent with the prototype of the other category (see Table 1).

Each trial contained a sample bug and a test bug. The sample bug was always one of the two prototypes used to produce the test bugs (Table 1). 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—horns, head, body, legs, or tail. The target feature of the test bug was covered by a mask. Two depictions of a feature were presented as choices with which to answer the inference question. One of these feature depictions was consistent with the sample bug, and the other was inconsistent with the sample bug (Figure 1). The participants indicated their choice of feature depiction by clicking the corresponding button.

Procedure and Design

At the beginning of the experiment, each participant was given one of the two instruction sheets, which divided the participants into the category condition or the feature condition, and the participants carried out the inference task, a filler task, and a similarity judgment task, in this order. Excerpts of the instructions from the two conditions are shown below:

Category condition. In this experiment, we are interested in the way you make judgments . . . Each bug belongs to two types—“monek” and “plaple.” These bugs are depicted with 5 different body parts—horns, head, body, legs, and tail, along with a tag, “monek” or “plaple” . . . Based on the samples shown on the left side of the screen, please answer each question as accurately as you can . . .

Feature condition. In this experiment, we are interested in the way you make judgments . . . Each bug is depicted with 6 different body parts—horns, head, body, legs, tail and wings. Because the wings of the bugs are folded on their back, we were not able to show them, so that we specified them with two names—“monek” and “plaple,” which roughly stand for two different shapes of wings . . . Based on the sample shown on the left side of the screen, please answer each question as accurately as you can . . .

The inference task consisted of 20 inference questions. For each trial, the 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 (Figures 1A and 1B). The sample stimuli were the prototype stimuli used to produce test stimuli (Table 1). Ten test stimuli were shown twice—once paired with the corresponding sample stimulus (matched trials; the sample and the test bugs had the same label) and once paired with a sample stimulus of the opposite category (mismatched trials; the sample and the test bugs had different labels). For example, Test 1a in Table 1 was shown twice, once with Sample A (a matched trial, because Test 1a and Sample A had the same label, *monek*) and once with Sample B (a mismatched trial, because Test 1a and Sample B had different labels, *monek* and *plaple*).

The dependent measure in this experiment was *consistency score* (the proportion of responses that selected the feature value consistent with the sample stimulus; Figures 1A and 1B). To assess the impact of labeling in the two instruction conditions further, we also calculated *polarity score* by subtracting the mean consistency score for the mismatched stimuli from that for the matched stimuli. To make the task more challenging, the 20 inference questions were randomly mixed with 20 other, unrelated inference questions. The order in which the trials were presented was determined randomly for each participant.

After the inference task, an unrelated filler task was given. This task took approximately 15 min. Following the filler task, the participants conducted a similarity judgment task, in which they indicated

the perceived similarity between the sample and the test stimuli on a 0–100 scale. The stimuli for the similarity task were analogous to those given in the inference task, except that the target questions, the mask, and the answer choices were removed (Figures 1C and 1D). The target features in the test stimuli were replaced with the feature consistent with the corresponding sample stimulus. This procedure was implemented in order to ensure that the participants’ similarity judgments corresponded to their inference performance. As in the inference task, the instructions in the category condition characterized the two labels as two *types of bugs*, and the instructions in the feature condition characterized the two labels as *two shapes of wings*. Other than this single point, the two conditions were identical.

The design of the experiment was 2 (instruction condition: category vs. feature—a between-subjects factor) × 2 (label match: matched vs. mismatched—a within-subjects factor) factorial. *Label match* refers to the two types of trials: matched (labels were matched) and mismatched (labels were mismatched).

Results and Discussion

As was predicted, characterizing the two arbitrary names as category labels modified the participants’ response patterns significantly (see Table 2).

A 2 (instruction condition: category vs. feature) × 2 (label match: matched vs. mismatched) ANOVA showed that there was a significant interaction effect between the two factors on the consistency score [$F(1,126) = 8.24$, $MS_e = 0.08$, $p < .01$, $\eta^2 = 0.06$]. The mean polarity scores, which were calculated by subtracting the consistency scores obtained in the mismatched trials from those obtained in the matched trials, were significantly larger in the category condition than in the feature condition [$t(126) = 2.82$, $p < .01$, $d = 1.26$]. There was a main effect of label match on the consistency score [see Table 2; $F(1,126) = 79.93$, $MS_e = 0.08$, $p < .01$, $\eta^2 = 0.38$]. The main effect of instruction condition was not significant ($F < 1$). To control the similarity and coverage factors, an item-based ANCOVA was performed with the similarity and coverage factors as covariates (see the Appendix for the procedure for estimating the coverage of stimuli).

Table 2
Mean Consistency Scores and Standard Deviations in Experiment 1

Condition	Match		Mismatch		Match – Mismatch	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Inference Task						
Category	.72*	.27	.30*	.26	.42**	.43
Feature	.62	.23	.40	.22	.22	.37
Similarity Judgment Task						
Category	.60	.15	.53	.13	.07	.08
Feature	.60	.16	.53	.16	.07	.10
Estimated Coverage						
Category	.34	.07	.19	.06	.14	.10
Feature	.35	.13	.23	.08	.13	.08

Note—To make a comparison between these values meaningful, the similarity judgment scores are divided by 100. For the procedure of estimating the coverage values, please see the Appendix. The asterisks show the results from *t* tests comparing two instruction conditions along each column. For example, in the column under “Match,” the mean score in the category condition ($M = .72$) was compared with that in the feature condition ($M = .62$). * $p < .05$. ** $p < .01$.

This analysis showed that even after controlling for these two factors, the effect of categorical labeling was substantial—an interaction effect between instruction condition and label match [$F(1,35) = 14.41$, $MS_e = 0.01$, $p < .01$].

These results are consistent with the following ideas: (1) Even when exactly the same materials are presented, drastically different inferential patterns can appear, depending on whether a label is described as a category label or as a feature label, and (2) this impact of category labels occurs even after the similarity and coverage factors (Slovan, 1993) are controlled. In Experiment 2, we examined whether category labels would influence the *decision processes*.

EXPERIMENT 2

Experiment 2 was designed to replicate Experiment 1 but also to further investigate the mechanisms of inductive inferences. Experiment 1 provided evidence that the participants were using category membership information differently from feature information. However, we do not know the exact role that category labels play in a *decision process*. By measuring how long and in what order a participant spent viewing a feature, we assumed that we could estimate how much that feature was guiding his/her decision process. In Experiment 2, the same conditions and stimuli as those in Experiment 1 were used, but the trials were presented using our mouse-tracking program.

Our predictions were as follows. First, if labels were used merely as a shortcut to make a response quickly, the overall response time for the participants in the category condition would be shorter than those in the feature condition. Second, if category labels were used to guide a decision, labels should be viewed longer and more frequently in the category condition than in the feature condition. Finally, if labels initiated an inference, labels should be viewed earlier in a trial in the category condition than in the feature condition.

Method

Participants

A total of 73 undergraduate students participated in the experiment. The data from 4 participants were removed because they did not conduct the experiment as suggested in the instructions.¹ To eliminate the outliers, we removed the data for the participants whose average viewing time for the stimulus areas exceeded 20 sec and whose average viewing times for the test stimulus area were below 1.2 sec.² We reasoned that at least 1.2 sec were needed to examine the stimulus areas and to make a judgment. Altogether, the data from 57 participants (category condition, $n = 28$; feature condition, $n = 29$) were analyzed. The assignment of the two conditions was made in an alternate order as the participants arrived at the laboratory. The participants were tested in groups ranging in size from 1 to 4.

Materials

The stimuli for this experiment were the same illustrations of fictional insects as those used in Experiment 1. To monitor mouse movement, a Java-based Restricted Focus Viewer program was used (Jansen et al., 2003).³ This program uses blurred images on the screen that can only be temporarily revealed by moving a small viewer window, using the mouse (Figure 2). Thus, to view a desired feature, the user must move the mouse to its location. This program records the movement of a mouse across the images, as well as the

amount of time spent at each location. The size of the viewer window for this experiment was 27×27 pixels.⁴ Our Restricted Focus Viewer program presented two stimuli (sample bug and test bug) that were blurred beyond visible recognition. As in Experiment 1, the target feature in the test bug was covered with a mask. Also blurred was the label (either *monek* or *plaple*) appearing above each of the bugs and inference questions attached to the test stimulus. To create the blurring effect, four different versions of 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 that 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 was inconsistent with the sample bug. The 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 the participants.

Table 3
Definitions of Dependent Variables and Other Terms

Areas of Interest (AoIs)

AoIs were defined on the computer screen. Each sample stimulus and test stimulus was divided into 12 different areas: label, horns, head, body, legs, and tail—in the sample and test stimuli (see Figure 3).

Fixation

If a cursor stayed at the same spot continuously for a time equal to or more than 50 msec, it was regarded as one fixation.

Fixation Time (ft)

The ft was the duration (measured in milliseconds) of one fixation. For example, if the cursor stayed at a particular position for 150 msec, the ft of that particular fixation was 150 msec.

Fixation Count (fc)

Any given fixation was counted as one fixation regardless of its ft. For example, there was a 100-msec fixation at one AoI and another 200-msec fixation at the same AoI some time later in the trial. This means that there were two fcs at that particular AoI.

Weighted Fixation Time (wft) and Weighted Fixation Count (wfc)

To assess the order of fixation, all occurrences of fixation in each trial were numbered in a descending order, and each fc and ft was weighted in accordance with its order of occurrence ($wfc = \text{weight} \times \text{fc}$; $wft = \text{weight} \times \text{ft}$). For example, the first fixation in a trial had a weight of N/N , the second fixation had a weight of $(N-1)/N$, and so on until the last fixation, which had a weight of $1/N$ (see Rehder & Hoffman, 2005a, for a similar procedure). Here, N stands for the number of total fcs in a trial.

Fixation Score and First Fixation

To identify when a label was viewed for the first time in each trial, we measured *first fixation score* of labels. Specifically, we (1) numbered every occurrence of fixation from 1 to N in each trial and (2) divided these numbers by the number of total fixations, creating a series of *fixation scores* (i.e., $\text{Fixation Scores}^{(N)} = \{1/N, 2/N, 3/N, \dots, N/N\}$, where N is a total number of fixations in a given trial). For example, assume that a total of 10 fixations occurred in a given trial in the following order: [Sample_Head, Test_Legs, Test_Tail, Sample_Label, Test_Tail, Test_Head, Sample_Label, Test_Label, Test_Legs, Test_Label]. The fixation scores of this example will be [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1], and the first fixation score of Sample_Label is 0.4 and that of Test_Label is 0.8. This statistic helps identify when a label was viewed in the course of decision making, independently of total viewing times of individual participants and trials. For example, the fixation that occurred at the 25th percentile of a viewing history of a trial will be .25, and the fixation that occurred at the 50th percentile of a viewing history will be .5. We recorded first fixation scores of labels and heads in all the trials.

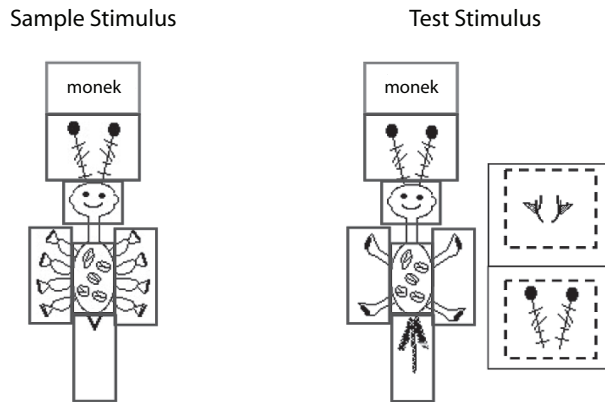


Figure 3. For the data analysis, the stimulus frame was divided into four different locations: the label of a sample stimulus (inside the upper rectangle on the left), the body parts of a sample stimulus (inside the lower rectangles on the left), the label of a test stimulus (inside the upper rectangle on the right), and the body parts of a test stimulus (inside the lower rectangles on the right). We also examined the viewing times of 12 individual components: two labels and 10 body parts in the sample and test stimuli.

Procedure and Design

At the beginning of the experiment, each participant was given one of the two instruction sheets. Next, all the participants viewed the same demonstration video clip, which showed how to use the Restricted Focus Viewer program. After the demonstration, the participants completed 20 trials at their own pace.

The design of Experiment 2 was a 2 (instruction condition: category vs. feature) × 2 (label match: matched vs. mismatched) factorial. For mouse tracking, viewing time data were analyzed using a 2 (instruction condition: category vs. feature) × 2 (stimulus area: sample vs. test) × 2 (label match: matched vs. mismatched) × 2 (mouse location: label vs. body part) factorial. Instruction condition was a between-subjects factor, and all the other factors were within-subjects manipulations. *Stimulus area* refers to the two depictions of stimuli—a sample stimulus and a test stimulus. *Mouse location* refers to where the participant was focusing the viewer window. Viewing times for body parts were calculated by averaging the viewing times for the five different body parts—horns, head, body, legs, and tail.

As in Experiment 1, the participants’ selection of a particular feature value was assessed by measuring the consistency score. For the mouse movement data, a stimulus frame was divided into 12 different areas of interest (AoIs; see Figure 3), and the following dependent variables were evaluated: fixation time, fixation count, and the order of fixation (see Table 3 for the definitions of these terms). These dependent measures and the analyses employed in this study were based on the eyetracking studies by Rehder and Hoffman (2005a, 2005b).

Results and Discussion

The individual trials whose viewing times were 0 msec were removed from the data analyses. Here, we defined *viewing time* as the time that a mouse stayed on sample and/or test stimulus areas. We also eliminated 1 participant in the feature condition because the responses of this participant deviated far from those of his/her group. This participant’s viewing times of labels exceeded 4.5 standard deviation units from the mean and 13 times more than the median of the condition. All the other responses were within 4.5 standard deviation units of the mean of

each condition. Because response time measures are particularly sensitive to outliers, we considered these procedures to be necessary (see Ratcliff, 1993; Wilcox, 1998).

Consistency score. The results from the consistency scores were in accord with those reported in Experiment 1 (see Table 4). As in Experiment 1, a 2 (label status) × 2 (instruction condition) ANOVA revealed a significant interaction between instruction condition and label match on the consistency score [$F(1,54) = 7.34, MS_e = 0.05, p < .01, \eta^2 = 0.12$]. To identify the locus of the interaction effect, we compared the mean polarity scores, which were calculated by subtracting the consistency scores obtained in the mismatched trials from those in the matched trials. The mean polarity score was larger in the category condition than in the feature condition [$t(54) = 2.71, p < .01, d = 0.72$]. The main effect of instructions was not significant ($F < 1$). Overall, matched stimuli received higher consistency scores than did mismatched stimuli [$F(1,54) = 28.65, MS_e = 0.05, p < .01, \eta^2 = 0.35$].

Mouse movement analyses. In our mouse movement analyses, we will report the results based on fixation count (fc) and fixation time (ft) and then the results based on weighted fixation count (wfc) and weighted fixation time (wft). We will report primarily interaction effects involving instruction condition.

As was predicted, labels were viewed longer and more frequently in the category condition than in the feature condition. There were significant interaction effects between instruction condition and mouse location [fc, $F(1,54) = 5.80, MS_e = 98.15, p < .05, \eta^2 = 0.1$; ft, $F(1,54) = 5.95, MS_e = 562,159.9, p < .05, \eta^2 = 0.1$]. The two labels were viewed more often and longer in the category condition than in the feature condition [fc, $t(54) = 2.23, p = .03, d = 0.60$; ft, $t(54) = 2.20, p = .03, d = 0.59$; see Table 5 and Figure 4A). Viewing times of the body parts did not differ between the two conditions [$ts(54) < 0.6, ps > .50, ds < 0.18$]. Two additional ANOVAs, which were applied separately to the data obtained from the sample stimulus and the test stimulus, suggested that categorical labeling influenced viewing times of the label in the test stimulus a great deal. Given the test stimulus, the interaction between instruction condition and mouse location was robust [fc, $F(1,54) = 7.65, MS_e = 28.86, p < .01, \eta^2 = 0.12$; ft, $F(1,54) = 6.33, MS_e = 159,885.22, p = .02, \eta^2 = 0.11$]. The interactions between the two factors were marginally significant in the sample stimulus [fc, $F(1,54) = 3.12, MS_e = 25.90, p = .08, \eta^2 = 0.55$; ft, $F(1,54) = 3.87, MS_e = 175,103.40, p = .05, \eta^2 = 0.07$; see Figure 4A].

Table 4
Mean Consistency Scores and Standard Deviations in Experiment 2

Condition	Match		Mismatch		Polarity Score (Match – Mismatch)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Category	.70	.28	.36*	.29	.34**	.37
Feature	.61	.23	.50	.27	.11	.26

Note—The asterisks show the results from *t* tests comparing two instruction conditions along each column. * $p < .05$. ** $p < .01$.

The three-way interaction between instruction condition, label match, and stimulus area was marginally significant for the fc measure [fc, $F(1,54) = 2.99$, $MS_e = 11.41$, $p = .09$, $\eta^2 = 0.05$], but not for the ft measure (ft, $F < 1$). All the other interactions related to instruction condition were not significant [$F_s(1,54) < 1.4$, $ps > .25$, $\eta^2 < 0.02$].

Categorical labeling per se neither prolonged nor reduced the total viewing time for the stimuli. A 2 (label match: matched vs. mismatched) \times 2 (instruction condition: category vs. feature) \times 2 (stimulus area: sample stimulus vs. test stimulus) ANOVA showed that modifying the instructions had no effect on the total viewing times for the trials [$F(1,54) = 1.53$, $MS_e = 2,802,036.14$, $p = .22$, $\eta^2 = 0.03$]. Thus, categorical labeling did not simply generate a quick-and-easy “short-cut” strategy. Rather, it led to a specific interaction between viewing locations and viewing times.

Now we will examine the impact of *viewing order*. If category labels were used to guide inference, they should be viewed earlier in a trial.

Viewing order. To investigate the effect of viewing order, the following measures were taken. First, we performed the same analyses as those described above, with the data set weighted with the order of viewing (see Table 3 for the procedure to obtain weighted data). Second, we identified the viewing “history” of the labels and investigated whether category labels were viewed earlier in a trial (the procedure of identifying *viewing history* will be described later).

Overall, the disparity between the two instruction conditions was intensified when viewing order was taken into account (see Figure 4B). The interaction effects between instruction condition and mouse location for fcs and fts were quite robust [wfc, $F(1,54) = 9.07$, $MS_e = 39.73$, $p < .01$, $\eta^2 = 0.14$; wft, $F(1,54) = 7.84$, $MS_e = 265,897.38$, $p < .01$, $\eta^2 = 0.13$; see Table 5]. Additional analyses applied separately to the test stimuli and the sample stimuli showed that this interaction was particularly conspicuous for the test stimuli, but it was also present in the sample stimuli [test stimulus, wfc, $F(1,54) = 11.56$, $MS_e = 9.38$, $p < .01$, $\eta^2 = 0.18$, and wft, $F(1,54) = 9.16$, $MS_e = 53,601.54$, $p < .01$, $\eta^2 = 0.15$; sample stimulus, wfc, $F(1,54) = 5.52$, $MS_e = 13.30$, $p = .02$, $\eta^2 = 0.093$, and wft, $F(1,54) = 4.50$, $MS_e = 110,448.77$, $p = .03$, $\eta^2 = 0.09$]. Why was the effect of category labels particularly

pronounced for the test stimuli, as compared with the sample stimuli? We do not have a clear answer to this question. It is possible that there is an asymmetrical relationship in the direction of comparison in feature inference (see A. B. Markman & Gentner, 1993).

In general, the two labels were viewed more often in the category condition than in the feature condition [fc, $t(54) = 2.63$, $p = .01$, $d = 0.72$; ft, $t(54) = 2.43$, $p = .02$, $d = 0.64$]. The viewing times for the body parts did not differ between the two instruction conditions [for all dependent measures, $t_s(54) < 1.34$, $ps > .18$, $ds < 0.37$]. There was no interaction between instruction condition and stimulus location [fc, $F(1,54) = 2.01$, $MS_e = 12.08$, $p = .16$, $\eta^2 = 0.04$; ft, $F(1,54) = 1.33$, $MS_e = 128,553.49$, $p = .25$, $\eta^2 = 0.02$].

These results support the idea that the labels were viewed earlier in the category condition than in the feature condition.

Viewing history: First fixations. Were labels viewed earlier in the category condition than in the feature condition? To answer this question, we compared the viewing histories of *labels* and *head* in individual trials (i.e., *first fixation score*; see Table 3 for the definition). It was shown by t tests that the first fixation of the label of a test stimulus occurred significantly earlier in the category condition than in the feature condition [$t(51) = 2.30$, $p = .03$, $d = 0.64$; see Table 6].⁵ In contrast, the head of the test stimulus was viewed earlier in the feature condition than in the category condition [$t(54) = 2.26$, $p = .03$, $d = 0.60$]. Given the sample stimulus, there was a tendency that the labels were viewed earlier in the category condition than in the feature condition [$t(53) = 1.78$, $p = .08$, $d = 0.48$]. The first fixation scores for the sample head did not differ in the two conditions [$t(54) = 0.90$, $p = .37$, $d = 0.24$].

Taken together, the results from this experiment indicate that category labels (1) were viewed more often and (2) were viewed earlier in a decision process, supporting the hypothesis that category labels, unlike other features, play a guiding role in inferential judgments.

GENERAL DISCUSSION

The goal of the present study was to investigate the influence of category labels on inductive inferences. We ex-

Table 5
Summary of Mouse Movement Data in Experiment 2

Condition	Fixation Count				Fixation Time			
	Label		Body Part		Label		Body Part	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Category	10.75*	6.82	10.72	5.65	855.32*	577.42	797.49	414.39
Feature	7.36	5.38	11.71	4.78	586.32	420.41	835.12	315.45
Condition	Weighted Fixation Count				Weighted Fixation Time			
	Label		Body Part		Label		Body Part	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Category	8.64*	4.76	6.24	2.66	695.51*	425.92	462.20	174.78
Feature	5.65	3.65	6.83	2.38	456.17	300.44	495.70	151.66

Note—The mean and standard deviation of fixation count and fixation time measures in Experiment 2. The numbers in the body part category represent the average of the five different body parts. The asterisks show the results from t tests comparing two instruction conditions along each column. * $p < .05$. ** $p < .01$.

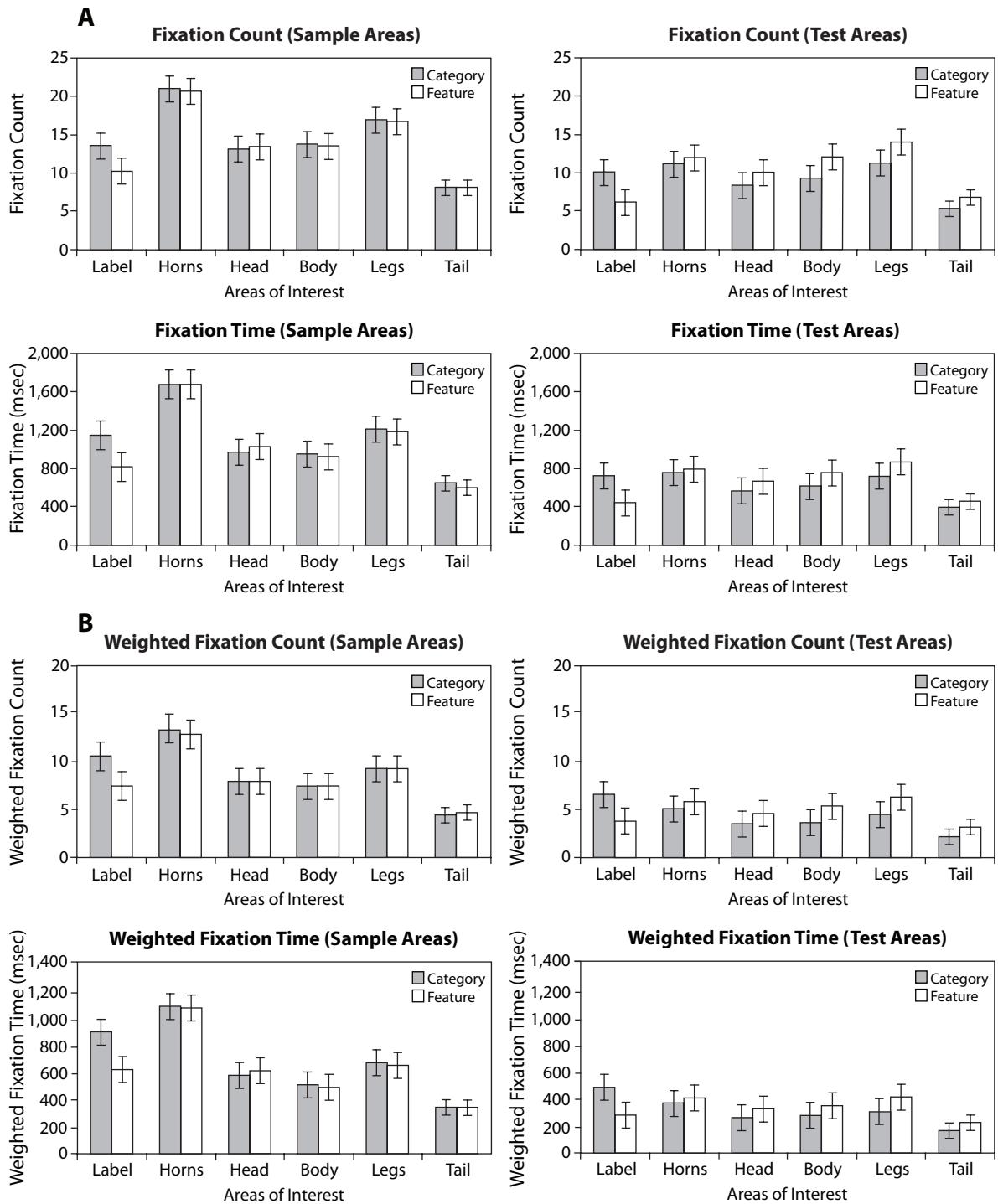


Figure 4. (A) Summary of fixation counts and fixation times obtained in Experiment 2. (B) Summary of weighted fixation counts and weighted fixation times collected in Experiment 2. The length of each error bar represents two units of weighted standard error obtained from the two instruction conditions. These bars correspond to two units of *t* score used for the *t* tests between the two instruction conditions.

explored this issue in the two experiments by manipulating the information the labels conveyed and then measuring the participants' inferences, as well as their mouse movement patterns. The results from Experiment 1 indicated that categorical information has a strong effect on inferences, even

after the similarity and coverage factors were being controlled. Trials in which the labels were congruent directed the participants to select a feature choice consistent with the sample stimulus, whereas incongruent label trials led the participants to select the feature choice inconsistent with

Table 6
Mean First Fixation Scores in Experiment 2

Condition	Sample Stimulus				Test Stimulus			
	Label		Head		Label		Head	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Category	.10	.06	.74	.08	.26*	.09	.49	.09
Feature	.13	.08	.76	.08	.33	.12	.43*	.11

Note—These numbers represent the means and standard deviations for the first fixation scores obtained over individual participants. The trials that had no fixation of labels or head were not included in this analysis. The participants who had fewer than three data points were not included in this analysis. Small numbers represent the fixations that occurred earlier in a trial. For example, .26 in the category condition of the test stimulus means that the first fixation of the label of a test stimulus occurred at the 26th percentile of the entire fixation history. The asterisks show the results from *t* tests comparing two instruction conditions along each column. * $p < .05$. ** $p < .01$.

the sample stimulus. This polarized response pattern was reduced when the labels were characterized as features.

The mouse movement data in Experiment 2 showed that category labels influence inductive inferences in a manner differently from feature labels. This was reflected in the viewing time and order data from the mouse movement program. Characterizing the labels with categorical information increased the proportion of viewing time spent on the labels in the test stimuli, as compared with the condition in which labels were characterized with feature information. Categorical labeling also influenced the order of viewing a stimulus. The labels were viewed earlier in a trial when the labels carried category information. This tendency was pronounced in the test stimuli.

These results are consistent with the ideas that (1) the presence of category labels (i.e., verbal labels that convey category membership information) redirects the way in which people approach the problem, (2) the category information has the effect of changing the strategy by which persons view the label and features, as well as manipulating the weight of the labels and features in their decision process, and (3) category labels and feature labels play different roles in inductive inference. Note that the feature labels in the present study were characterized as representing the *shapes* of wings. In a separate study, we also introduced feature labels representing *different islands*, *different kinds of food*, and *different kinds of disease* that bugs live, eat, or carry and compared them with category labels. Even with these manipulations, we observed results analogous to those found in the present study (Yamauchi & Yu, 2005, 2007).

It is difficult to interpret these results with similarity-based theories of inductive inference alone (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Pothos, 2005; Sloman, 1993, 1998; Sloutsky, 2003; Sloutsky & Fisher, 2004). According to the similarity-based approach, labels are only salient features. If this were the case, we would not expect label-viewing times and orders to differ between the category and the feature conditions. The two conditions had the same verbal labels and the same stimuli. Thus, the disparity between the two conditions should come only from the *category information* presented in the instructions, not from the presence/absence of verbal

labels per se, as has been examined in previous studies (Gelman & Markman, 1986; Sloutsky & Fisher, 2004).

It is possible that the category information given in the instructions made the verbal labels *attentionally salient*, leading the participants to view labels earlier and more often. The viewing order data in Experiment 2 show that the labels conveying category information were viewed earlier, as compared with labels associated with feature information, implying that the category labels gave a goal-driven preparatory clue. All the stimuli in our experiments had the same labels, the participants received the same stimulus frames, and the manipulation was made solely in instructions. Thus, the attentional salience observed in the two experiments was unlikely to be stimulus driven.

Why does categorical labeling change people's inductive strategy? We speculate that there are two possibilities. A first possibility concerns the effect of communication. In order for linguistic communication to be effective, speakers and listeners must establish an agreement as to what a particular word stands for. For example, some object is called *booklet*, rather than *pamphlet*, chiefly because that particular reference was established earlier between communicators (Malt & Sloman, 2004; Malt, Sloman, & Gennari, 2003; Markman & Makin, 1998). This communicative "pact" may become particularly strong when labels represent a category of objects, rather than a feature of objects. A second possibility concerns ecologically driven reasoning heuristics (Gigerenzer, Todd, & the ABC Research Group, 1999). Natural categories are clustered by family resemblance (Rosch & Mervis, 1975). In many basic-level categories, no single feature defines categorical boundaries. In this setting, using category membership as a default reasoning strategy might be a robust and reasonably satisfying means of induction. It is possible that our adult participants learned such a reasoning strategy as they experienced numerous categories in their lives (Sloutsky, 2003; Sloutsky & Fisher, 2004). It is also possible that the category-based reasoning strategy, which may be related to *psychological essentialism* (Medin & Ortony, 1989), is ingrained through evolution as an ecologically rational decision process (Gelman, 2003). Future studies should investigate this issue.

CONCLUSION

This experiment has shed light on how people use category labels and feature labels to make inductive inferences; it has also demonstrated a new method for studying inductive-reasoning research. Although we cannot be totally sure what participants are thinking when they view stimuli, the combination of decision responses and decision process measures inferred from mouse movement patterns can point to the idea that category membership information plays a distinct role in feature inference and separates itself from other features.

AUTHOR NOTE

This article is an extension of the paper "Feature Inference: Tracking Mouse Movement," presented at the 27th Annual Meeting of the

Cognitive Science Society. This research was supported by a Glasscock Center Faculty Fellow Award, and a Developmental Grant by the Mexican American and U. S. Latino Research Center, Texas A&M University, which were given to the first author. The authors thank Art Markman and Chase Chick for their comments. Please address all correspondence to T. Yamauchi, Department of Psychology, Texas A&M University, Mail Stop 4235, College Station, TX 77843 (e-mail: tya@psyc.tamu.edu).

REFERENCES

- ALLPORT, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.
- ANDERSON, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- CORNEILLE, O., KLEIN, O., LAMBERT, S., & JUDD, C. M. (2002). On the role of familiarity with units of measurement in categorical accentuation: Tajfel and Wilkes (1963) revised and replicated. *Psychological Science*, **13**, 380-383.
- DUCKITT, J. (1992). Psychology and prejudice: A historical analysis and integrative framework. *American Psychologist*, **47**, 1182-1193.
- GELMAN, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford: Oxford University Press.
- GELMAN, S. A., & COLEY, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inference in 2-year-old children. *Developmental Psychology*, **26**, 796-804.
- GELMAN, S. A., & HEYMAN, G. D. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological Science*, **10**, 489-493.
- GELMAN, S. A., & MARKMAN, E. M. (1986). Categories and induction in young children. *Cognition*, **23**, 183-209.
- GIGERENZER, G., TODD, P. M., & THE ABC RESEARCH GROUP (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
- HAMILTON, D. L., & SHERMAN, J. W. (1994). Stereotypes. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 2, pp. 1-68). Hillsdale, NJ: Erlbaum.
- HEIT, E., & RUBINSTEIN, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 411-422.
- JANSEN, A. R., BLACKWELL, A. F., & MARRIOTT, K. (2003). A tool for tracking visual attention: The Restricted Focus Viewer. *Behavior Research Methods, Instruments, & Computers*, **35**, 57-69.
- KAHNEMAN, D., & TVERSKY, A. (1973). On the psychology of prediction. *Psychological Review*, **80**, 237-251.
- KASHIMA, Y., WOOLCOCK, J., & KASHIMA, E. S. (2000). Group impressions as dynamic configurations: The tensor product model of group impression formation and change. *Psychological Review*, **107**, 914-942.
- KRUSCHKE, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.
- KUNDA, Z., MILLER, D. T., & CLAIRE, T. (1990). Combining social categories: The role of causal reasoning. *Cognitive Science*, **14**, 551-577.
- KUNDA, Z., & THAGARD, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, **103**, 284-308.
- LOVE, B. C., MEDIN, D. L., & GURECKIS, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, **111**, 309-332.
- MALT, B. C., & SLOMAN, S. A. (2004). Conversation and convention: Enduring influences on name choice for common objects. *Memory & Cognition*, **32**, 1346-1354.
- MALT, B. C., SLOMAN, S. A., & GENNARI, S. P. (2003). Universality and language specificity in object naming. *Journal of Memory & Language*, **49**, 20-42.
- MARKMAN, A. B., & GENTNER, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, **25**, 431-467.
- MARKMAN, A. B., & MAKIN, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General*, **127**, 331-354.
- MARKMAN, A. B., & ROSS, B. H. (2003). Category use and category learning. *Psychological Bulletin*, **129**, 592-613.
- MARKMAN, E. M. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.
- MEDIN, D. L., COLEY, J. D., STORMS, G., & HAYES, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, **10**, 517-532.
- MEDIN, D. L., & ORTONY, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179-195). Cambridge: Cambridge University Press.
- MURPHY, G. L., & ROSS, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, **27**, 148-193.
- NOSOFSKY, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.
- OSHERSON, D. N., SMITH, E. E., WILKIE, O., LOPEZ, A., & SHAFIR, E. (1990). Category-based induction. *Psychological Review*, **97**, 185-200.
- POTHOS, E. M. (2005). The rules versus similarity distinction. *Behavioral & Brain Sciences*, **28**, 1-49.
- RATCLIFF, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, **114**, 510-532.
- REHDER, B., & HOFFMAN, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, **51**, 1-41.
- REHDER, B., & HOFFMAN, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 811-829.
- ROSCH, E., & MERVIS, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.
- ROSS, B. H., GELMAN, S. G., & ROSENGREN, K. S. (2005). Children's category-based inferences affect classification. *British Journal of Developmental Psychology*, **23**, 1-24.
- SHAFIR, E. B., SMITH, E. E., & OSHERSON, D. N. (1990). Typicality and reasoning fallacies. *Memory & Cognition*, **18**, 229-239.
- SLOMAN, S. A. (1993). Feature-based induction. *Cognitive Psychology*, **25**, 231-280.
- SLOMAN, S. A. (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**, 246-251.
- SLOUTSKY, V. M., & FISHER, A. V. (2004). Induction and categorization in young children. *Journal of Experimental Psychology: General*, **133**, 166-188.
- STANGOR, C. (2000). Volume overview. In C. Stangor (Ed.), *Stereotypes and prejudice: Essential readings* (pp. 1-16). Philadelphia: Psychology Press.
- TAJFEL, H., & WILKES, A. L. (1963). Classification and quantitative judgment. *British Journal of Psychology*, **54**, 101-114.
- TVERSKY, A., & KAHNEMAN, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, **90**, 293-315.
- WILCOX, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, **53**, 300-314.
- YAMAUCHI, T. (2005). Labeling bias and categorical induction: Generative aspects of category information. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 538-553.
- YAMAUCHI, T., & MARKMAN, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 776-795.
- YAMAUCHI, T., & YU, N. Y. (2005). Categories and feature inferences: Category membership and a reasoning bias. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 2404-2409). Mahwah, NJ: Erlbaum.
- YAMAUCHI, T., & YU, N. Y. (2007). *Categories versus feature labels: Category labels polarize inferential predictions*. Manuscript in preparation.

NOTES

1. These participants responded almost exclusively with one of the two designated buttons (for all cases, $p < .001$; binomial distribution test).
2. Following the recommendation of Ratcliff (1993), we reanalyzed the mouse movement data with several different cutoff points—removing the

data from the participants whose viewing times exceeded 18, 19, 20, 21, or 22 sec, and removing the data from the participants whose average viewing times of the test stimuli were less than 400, 600, or 800 msec, all the way up to 2,000 msec. The overall results of these additional analyses were consistent with those described in the Results and Discussion section.

3. Restricted Viewer Program can be downloaded at www.csse.monash.edu.au/projects/RfV/.

4. The computer screen had $1,027 \times 768$ pixels, and the viewer window measured approximately 0.85×0.85 cm². The labels were placed about 1.2 cm above the top of the horns, so that the participants were unable to see a label and other body parts (e.g., horns) simultaneously.

5. The t tests described here were calculated over the average first fixation scores obtained from individual participants. The participants who had fewer than three data points were not included in these t tests.

APPENDIX

Estimating the Coverage Factor: An Analysis of the Sloman (1993) Model

Sloman's (1993) feature-based model provides simple yet elegant explanations for most of the psychological phenomena examined by Osherson et al. (1990). For this reason, we adopted the Sloman model to estimate the coverage factor of induction. The model explains the strength of a conclusion, such as *lions have disease X*, given a premise, *zebras have disease X*, by Equation A1:

(*P*: Premise) Zebras have disease X.

(*C*: Conclusion) Lions have disease X.

$$a(C|P) = \frac{F(P) \cdot F(C)}{|F(C)|^2} = \frac{|F(P)| [F(P) \cdot F(C)]}{|F(P)||F(C)||F(C)|} = \frac{|F(P)|}{|F(C)|} \text{sim}(P, C) \quad (\text{A1})$$

where

$$\text{sim}(P, C) = \frac{F(P) \cdot F(C)}{|F(P)||F(C)|}, \quad F(P) \cdot F(C) = \sum_{i=1}^n p_i c_i, \quad \text{and} \quad |F(P)| = \sqrt{\sum_{i=1}^n p_i^2}.$$

$a(C|P)$ stands for the strength of conclusion C , given premise P . $F(P)$ and $F(C)$ are vectors representing an item in premise P (i.e., *zebras*) and in conclusion C (i.e., *lions*), respectively. Each element of these vectors specifies a feature value of an item (e.g., having four legs).

In our experiments, we treated a sample stimulus in a stimulus frame as a premise and a test stimulus as a conclusion (see Figure 1) and translated Equation A1 into Equations A2A and A2B:

$$a(T_i^k | S_i^k) = \frac{F(S_i^k)}{F(T_i^k)} \text{sim}(S_i^k, T_i^k) \quad (\text{A2A})$$

and

$$a(T_i^{k'} | S_i^{k'}) = \frac{F(S_i^{k'})}{F(T_i^{k'})} \text{sim}(S_i^{k'}, T_i^{k'}). \quad (\text{A2B})$$

Equations A2A and A2B correspond to trials in the category condition and in the feature condition, respectively. Each of S_i^k , T_i^k , $S_i^{k'}$, and $T_i^{k'}$ stands for vectors of a sample stimulus or a test stimulus given in the category condition or in the feature condition. The superscripts k and k' represent the values of the labels (*monek* or *plaple*) that are given in the category condition (k) or in the feature condition (k') ($\{k, l, k', l' | k, l, k', l' \in (\text{Monek} \cup \text{Plaple}), k \neq l, k' \neq l'\}$), and subscript i stands for the value of the target feature ($\{i, j, | i, j \in (\text{Value}_0 \cup \text{Value}_1), i | j\}$). We assume that the consistency score (i.e., $P(T_i^k | S_i^k)$) and the argument strength [$a(T_i^k | S_i^k)$] are monotonically related by some logistic regression function:

$$P(T_i^k | S_i^k) = \frac{\exp\{\beta_0 + \beta_1 a(T_i^k | S_i^k)\}}{1 + \exp\{\beta_0 + \beta_1 a(T_i^k | S_i^k)\}} = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 a(T_i^k | S_i^k))\}} \quad (\text{A3A})$$

and

$$P(T_i^{k'} | S_i^{k'}) = \frac{\exp\{\beta_0 + \beta_1 a(T_i^{k'} | S_i^{k'})\}}{1 + \exp\{\beta_0 + \beta_1 a(T_i^{k'} | S_i^{k'})\}} = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 a(T_i^{k'} | S_i^{k'}))\}} \quad (\text{A3B})$$

The same assumption can be applied to the polarity score by calculating the consistency scores for matched and mismatched stimuli separately and then obtaining the difference between the two.

The present analysis is concerned with whether or not the similarity and coverage factors alone can account for the effect of category labels. Thus, the two parameters of the logistic regression functions (β_0 and β_1) are fixed. Note that the Sloman model reveals that the consistency score [i.e., $P(T_i^k | S_i^k)$] is monotonically related to the two variables (similarity and coverage), and no inductive differences are expected to arise whether labels carry category information or feature information, as long as the influence of these two components is equivalent in the two instruction conditions.

To estimate the coverage factor, two independent pilot studies were conducted. In Pilot Study 1 ($N = 70$), all verbal labels were removed from the test stimuli, and the participants answered inference questions when only the sample stimuli had verbal labels. In Pilot Study 2, this condition was reversed. All the verbal labels were removed from the sample stimuli, and the participants answered inference questions when only the test stimuli carried verbal labels. Except for these two points, the materials, procedure, and design of Pilot Studies 1 and 2 were identical to those described in Experiment 1. The products of inference scores obtained in these two pilot studies were assumed as estimates of the coverage factor (e.g., $|F(P)| / |F(C)|$ in Equation A1).

Equations A4A and A4B show an estimated consistency score in Pilot Study 1 given in the category condition (A4A) or in the feature condition (A4B):

$$a(T_i^\phi | S_i^k) = \frac{|F(S_i^k)|}{|F(T_i^\phi)|} \text{sim}(S_i^k, T_i^\phi) \quad (\text{A4A})$$

and

$$a(T_i^\phi | S_i^{k'}) = \frac{|F(S_i^{k'})|}{|F(T_i^\phi)|} \text{sim}(S_i^{k'}, T_i^\phi) \quad (\text{A4B})$$

Equations A5A and A5B show an estimated consistency score in Pilot Study 2 given in the category condition (A5A) or in an attribute condition (A5B):

$$a(T_i^k | S_i^\phi) = \frac{|F(S_i^\phi)|}{|F(T_i^k)|} \text{sim}(S_i^\phi, T_i^k) \quad (\text{A5A})$$

and

$$a(T_i^{k'} | S_i^\phi) = \frac{|F(S_i^\phi)|}{|F(T_i^{k'})|} \text{sim}(S_i^\phi, T_i^{k'}) \quad (\text{A5B})$$

In Pilot Study 1, the test stimuli had no label (T_i^ϕ), and in Pilot Study 2, the sample stimuli had no label (S_i^ϕ).

By multiplying Equations A4A and A5A and Equations A4B and A5B, we obtain Equations A6A and A6B, respectively:

$$\begin{aligned} \frac{|F(S_i^k)|}{|F(T_i^\phi)|} \text{sim}(S_i^k, T_i^\phi) \times \frac{|F(S_i^\phi)|}{|F(T_i^k)|} \text{sim}(S_i^\phi, T_i^k) &= \frac{|F(S_i^k)|}{|F(T_i^\phi)|} \times \frac{|F(S_i^\phi)|}{|F(T_i^k)|} \times \text{sim}(S_i^k, T_i^\phi) \times \text{sim}(S_i^\phi, T_i^k) \\ &= \frac{|F(S_i^k)|}{|F(T_i^k)|} \times u \end{aligned} \quad (\text{A6A})$$

and

$$\begin{aligned} \frac{|F(S_i^{k'})|}{|F(T_i^\phi)|} \text{sim}(S_i^{k'}, T_i^\phi) \times \frac{|F(S_i^\phi)|}{|F(T_i^{k'})|} \text{sim}(S_i^\phi, T_i^{k'}) &= \frac{|F(S_i^{k'})|}{|F(T_i^{k'})|} \times \frac{|F(S_i^\phi)|}{|F(T_i^\phi)|} \times \text{sim}(S_i^{k'}, T_i^\phi) \times \text{sim}(S_i^\phi, T_i^{k'}) \\ &= \frac{|F(S_i^{k'})|}{|F(T_i^{k'})|} \times u \end{aligned} \quad (\text{A6B})$$

where

$$u = \frac{|F(S_i^\phi)|}{|F(T_i^\phi)|} \times \text{sim}(S_i^k, T_i^\phi) \times (S_i^\phi, T_i^k),$$

and

$$u = \frac{|F(S_i^\phi)|}{|F(T_i^\phi)|} \times \text{sim}(S_i^{k'}, T_i^\phi) \times (S_i^\phi, T_i^{k'}),$$

Note that u and u' are identical, except for their similarity components. Because two other pilot studies ($N = 95$) estimating the similarity components of u and u' [$\text{sim}(S_i^k, T_i^\phi) \times \text{sim}(S_i^\phi, T_i^k)$] and [$\text{sim}(S_i^{k'}, T_i^\phi) \times \text{sim}(S_i^\phi, T_i^{k'})$] showed no statistical difference between the two instruction conditions [item-based t tests, $t(38) = 0.91$, $p = .36$, $d = 0.30$], we assumed $u \cong u'$. Accordingly, products of inference scores obtained in the two pilot studies (Equations 6A and 6B) were used as the estimates of the coverage factor.