Journal of Experimental Psychology: Applied
2010, Vol. 16, No. 4, 361–377

1076-898X/10/$12.00 DOI: 10.1037/a0021610

Learning and Retention Through Predictive Inference and Classification

Yasuaki Sakamoto
Stevens Institute of Technology

Bradley C. Love
The University of Texas at Austin

Work in category learning addresses how humans acquire knowledge and, thus, should inform classroom practices. In two experiments, we apply and evaluate intuitions garnered from laboratory-based research in category learning to learning tasks situated in an educational context. In Experiment 1, learning through predictive inference and classification were compared for fifth-grade students using class-related materials. Making inferences about properties of category members and receiving feedback led to the acquisition of both queried (i.e., tested) properties and nonqueried properties that were correlated with a queried property (e.g., even if not queried, students learned about a species' habitat because it correlated with a queried property, like the species' size). In contrast, classifying items according to their species and receiving feedback led to knowledge of only the property most diagnostic of category membership. After multiple-day delay, the fifth-graders who learned through inference selectively retained information about the queried properties, and the fifth-graders who learned through classification retained information about the diagnostic property, indicating a role for explicit evaluation in establishing memories. Overall, inference learning resulted in fewer errors, better retention, and more liking of the categories than did classification learning. Experiment 2 revealed that querying a property only a few times was enough to manifest the full benefits of inference learning in undergraduate students. These results suggest that classroom teaching should emphasize reasoning from the category to multiple properties rather than from a set of properties to the category.

Keywords: category learning, classroom learning, retention, classification, predictive inference

Work in category learning addresses how humans acquire knowledge, and thus it should influence and be influenced by classroom learning. Unfortunately, the link between these two domains is not as solid as one may expect. In the area of memory research, insights into basic memory processes garnered from laboratory studies have been applied to develop teaching instructions (Pavlik & Anderson, 2008) and to improve classroom performance for undergraduates and younger population (Metcalfe & Kornell, 2007). Similarly, progress in category learning research should impact classroom instruction. Conversely, considering the demands of the classroom setting highlights additional factors, such as efficient acquisition and long-term retention of material, which are crucial to developing comprehensive theories but are often neglected in category learning research.

One purpose of the current work was to transition the basic research in category learning with undergraduates to primary school students using class-related materials. In Experiment 1, fifth-grade students' category knowledge resulting from classification learning was compared with that resulting from inference learning. Classification and inference learning are two aspects of category acquisition. In classification learning, the learner acquires category knowledge by classifying items. For example, a child may classify an animal as a bird using its properties, such as small and flying. The child may be told that the animal is indeed a bird, strengthening the hypothesis about which properties predict the bird category, or that the animal is a mammal, leading to a modification of the hypothesis. In inference learning, the learner acquires category knowledge by inferring properties of category members. For example, a child may infer characteristics of birds, and a medical student may infer symptoms of patients with a certain disease. Based on the corrective feedback, they learn the category-property associations.

Existing theories of category learning do not specify what information is retained after inference and classification learning. Moreover, the processes underlying inference learning are not as well understood as those underlying classification learning. In Experiment 2, the roles direct queries play in inference learning are examined with undergraduate students. The results from the present work will extend the existing theories of how people acquire and retain category knowledge, and help the establishment of guidelines on the delivery of educational materials. These are the theoretical and practical contributions of the current work. Before presenting our predictions and results, we describe the classification and inference learning tasks in laboratory studies, and review a subset of relevant work. We conclude by discussing the implications of our results for theories of category learning and instructional practice.

Yasuaki Sakamoto, Howe School of Technology Management, Stevens Institute of Technology; Bradley C. Love, Department of Psychology, The University of Texas at Austin.

This work was supported by AFOSR FA9550-10-1–0268 and NSF CAREER #0349101 to B.C. L. We thank Momoko Sato and Tyler Davis for their help in collecting data. A portion of Experiment 1 was presented at the 28th Annual Conference of the Cognitive Science Society (Sakamoto & Love, 2006b). Experiment 2 was presented at the 31st Annual Conference of the Cognitive Science Society (Sakamoto & Love, 2009).

Correspondence concerning this article should be addressed to Yasuaki Sakamoto, Stevens Institute of Technology, Castle Point on Hudson, Hoboken, NJ 07030. E-mail: ysakamot@stevens.edu

Classification and Inference Learning Tasks

Classification and inference learning are two common tasks in category learning experiments. In this section, we detail these two tasks, and relate them to classroom learning.

Classification Learning Task

In a classification learning trial, the learner is presented with an item's properties, is asked to predict the category membership of the item, and then receives corrective feedback after responding. Figure 1A shows a

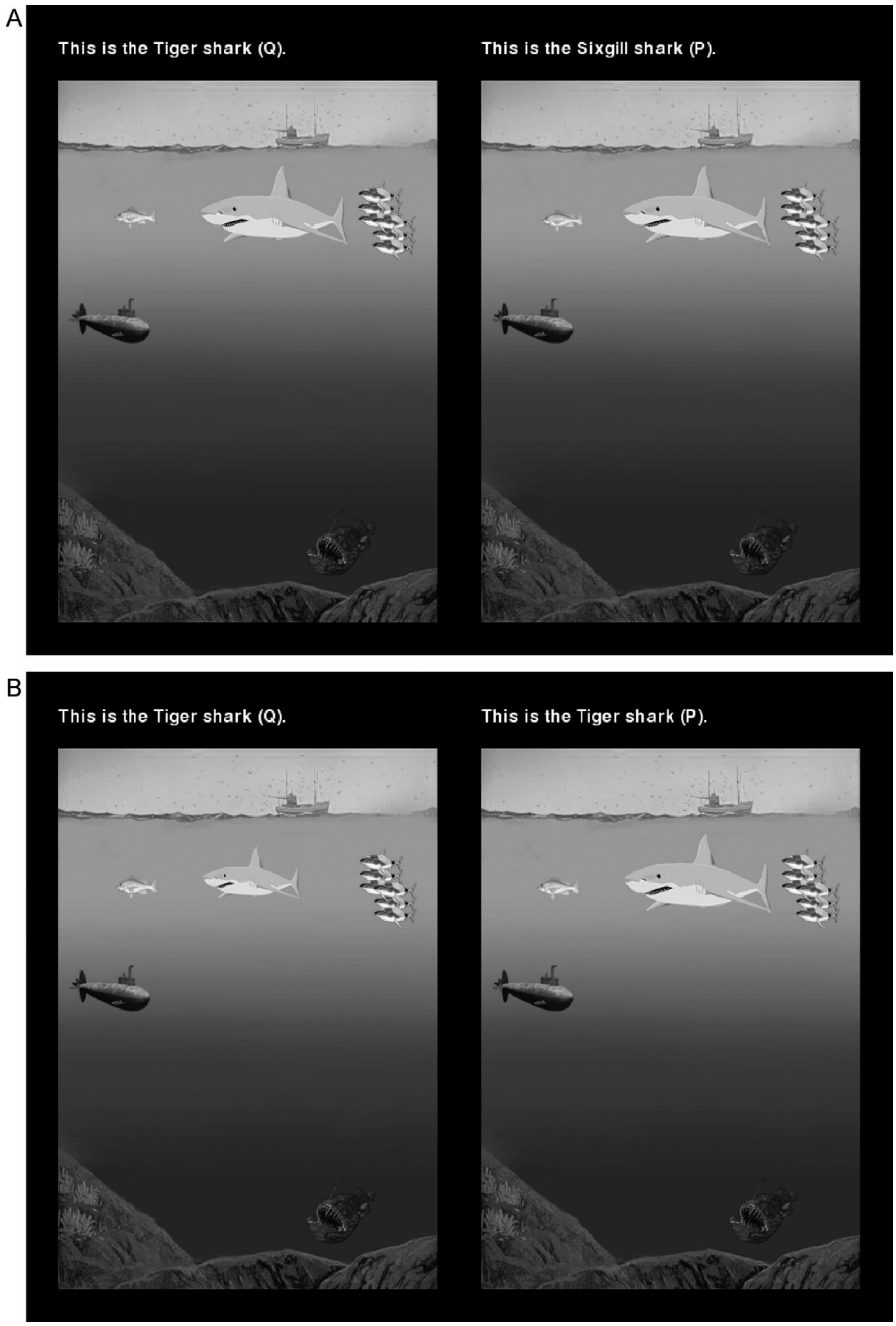

Figure 1. [A] A classification learning trial from Experiment 1 is shown. The learner pressed the Q or the P key to indicate whether the left or the right side correctly described the shark, respectively. The two descriptions differed only on the category label (Tiger vs. Sixgill shark): the learner predicted the category label. [B] An inference learning trial from Experiment 1 is shown. The learner guessed whether the left or the right side correctly described the shark as in the classification learning. The two descriptions differed only on the values of the queried dimension (smaller vs. larger body size): the learner predicted the property of a given shark.

snapshot of a classification learning trial from the current experiment. In this example, the learner guesses whether the left or right side describes the shark correctly. The left side of Figure 1A conveys the same information as the right side except for the category labels. On this trial, the learner predicts whether the given shark is a member of the Tiger or Sixgill shark. After responding, the learner receives corrective feedback, consisting of the correct category label and all properties of the item (i.e., the side that correctly describes the shark). Over the course of training, the learner completes a series of classification learning trials.

Participants in classification learning experiments develop category knowledge by classifying items into contrasting categories. Classification learning is important as the ability to retrieve category knowledge from memory and make predictive inference, such as birds have wings and fly, depends on first classifying the animal as birds. However, people also learn about categories during inference episodes. Thus, researchers have focused on inference learning, realizing that although work on classification learning has advanced the development of theories about people's classification behavior, these theories do not generalize to other aspects of category learning (see Markman & Ross, 2003 for a review).

Inference Learning Task

In an inference learning trial, the learner is presented with a subset of an item's properties along with the item's category membership, is queried about the value of an unknown property, and then receives corrective feedback after responding. Figure 1B displays a snapshot of an inference learning trial from the present experiment. In this example, the learner guesses whether the left or right side describes the shark correctly, as in the previous classification learning example. The left side of Figure 1B conveys the same information as the right side except for the queried dimension, size. The category label, Tiger shark, appears on each side of the screen, indicating that this trial is about the Tiger shark. The learner predicts whether the Tiger shark is small or large, and then, as in classification learning, is shown the side that correctly describes the shark. Over the course of training, the learner completes a series of inference learning trials with the queried property varying across trials, unlike classification learning in which the category label is always queried.

The learner in inference learning develops category knowledge through inferring multiple properties of stimulus items. Besides this difference, the two learning tasks are similar. In fact, the category label and all properties of items are presented to the learners in both learning tasks.

Classification and Inference Learning in Classroom Settings

Of course, primary school students acquire knowledge through various learning experiences inside and outside of the classroom. Nevertheless, parents and teachers do frequently ask classification-type and inference-type questions, although such questions may not be the main focus of teaching. For instance, parents often introduce concepts by pointing to an object and labeling it (e.g., Callanan, 1985; Ninio, 1980; Ninio & Bruner, 1978), essentially practicing classification. In the classroom, teachers use flashcards with pictures to teach students category memberships of objects (e.g., Is this a mammal or a bird?) and properties of different objects (e.g., Do mammals lay eggs or give birth to live offspring?). Moreover, many tests and assignments involve questions in the forms of classification and inference. Educators believe that exercises involving classification and inference can strengthen students' knowledge, and recommend the use of these exercises coupled with other activities (Rule, 2007). However, the benefits of classification and inference training to primary school students' learning are unclear. Applying inference and classification tasks to this population in a classroom setting has an important educational implication.

What Is Known About Classification and Inference Learning

In this section, we review how seemingly minor differences between the inference and classification training procedures lead to large differences in people's category knowledge.

Organization of Category Knowledge

Humans adapt their category representations to meet the demands of the task (see Love, 2005 for a review). For example, whereas maintenance workers organize the tree category around weediness, landscapers organize the tree category around both height and weediness, consistent with the different goals of these tree experts (Lynch, Coley, & Medin, 2000). Like different types of tree experts, the learner has different goals in classification and inference learning tasks. Whereas classification learning stresses the discrimination of members from different categories, inference learning stresses the discovery of the internal structure of each category. Consistent with the different focuses of the two tasks, whereas learners in classification learning acquire information that is diagnostic in distinguishing different categories during learning, those in inference learning acquire both diagnostic information and other information that frequently occurs within each category (e.g., Chin-Parker & Ross, 2004; Yamauchi, Love, & Markman, 2002; Yamauchi & Markman, 1998).

One question that the previous studies do not directly answer is whether learners retain more category information after inference or classification learning. Learners' focus on diagnostic information under classification learning does not necessarily mean that they have no memory trace for other information. They may retain nondiagnostic information but do not use it when making decisions. In Experiment 1, learners' retention of category information after classification and inference learning are compared.

Ease of Learning

The different focuses of classification and inference learning tasks also lead to different kinds of categories being better suited to these two tasks (Yamauchi & Markman, 1998; Yamauchi et al., 2002). Most real world categories follow family resemblance structures (see Rosch & Mervis, 1975), in which properties co-occur but no single property is common to all members of a category. For such categories, inference learning results in fewer learning errors and faster mastery of the categories than classification learning (Yamauchi & Markman, 1998). Many categories primary school students learn in the classroom, such as different

animals, follow family resemblance structures (Murphy, 2002). Thus classroom learning should be easier through inference than classification. The ease of learning is measured by error rate during learning in the current work.

Efficient acquisition of material is important in classroom learning. However, it may have a negative side effect. According to the error-driven learning account, errors mediate memory storage (Rescorla & Wagner, 1972) by leading to greater focus on error-producing items (e.g., Mackintosh, 1975). Then, classification learning, which results in more errors than inference learning, may improve students' memory for category information. Whether inference or classification learning is more beneficial for classroom learning is examined in Experiment 1.

Processes Underlying Classification and Inference Learning

Theories of category learning agree that learners shift attention to one or more diagnostic dimensions under classification learning (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004). In family resemblance categories, focusing on a diagnostic property to make classification decisions, such as flying for birds, leads to errors on exception items that violate the regularity, such as bats. It has been proposed that the learner's category representation under classification learning includes diagnostic information plus information about a few exception items (Love et al., 2004; Nosofsky, Palmeri, & McKinley, 1994). This representation is useful for classifying objects but not for other tasks, such as inferring properties of objects.

How people process information during inference learning is not well understood. Love et al. (2004) has proposed that learners under inference learning represent a category with one or more prototypes, or bundle of related feature dimensions, such as has fins, swims, and breathes underwater. According to this prototype account, learning one correlation (e.g., has fins co-occurs with swims) necessitates learning all of the correlations (e.g., has fins, swims, and breathes underwater all co-occur) when properties are interrelated. This account predicts that learners in inference learning will learn about a stimulus dimension that is not directly queried when this dimension correlates with stimulus dimensions that are queried. An alternative account holds that learners store a set of unrelated category-property mapping rules, such as "if fish, then has fins; if fish, then swims; if fish, then breathes underwater," that are sufficient to complete the inference task (Johansen & Kruschke, 2005). According to this rule account, when all stimulus dimensions are queried, learners will appear as if they have formed a prototype, but are in reality learning a set of rules. This account predicts that information about nonqueried dimensions will not be acquired in inference learning. We test the predictions of these two accounts.

Outstanding Theoretical and Applied Issues

One open theoretical issue is the nature of long-term retention in inference and classification learning. It is not entirely clear whether classification or inference learning results in the retention of more category information. Previous work has shown that testing on material can be more beneficial to establishing memories than additional study (e.g., Nungester & Duchastel, 1982; see Roediger & Karpicke, 2006 for a review). Similarly, queries in inference learning, which directly test knowledge of category properties, may play an important role in strengthening memories. In contrast, classification learning, in which learners need to discover which properties are associated with the categories by themselves, may not be beneficial for retention of category knowledge. Better understanding of the nature of retention has obvious importance in education and other everyday activities. Nevertheless, the significance of retention to memory performance has been underappreciated by cognitive psychologists (Wixted, 2005) and grossly neglected by category learning researchers. For instance, studies in category learning that examine memory often impose delays of only a few minutes (e.g., Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004). The present work is unique in that retention is measured over multiple-day delay. In addition to addressing the nature of retention, we extend research in category learning to fifth-graders learning about class-related materials in Experiment 1.

Another unexplored theoretical issue is how people process information under inference learning. We examine whether learners exclusively process queried dimensions during inference learning (Johansen & Kruschke, 2005) or they also process nonqueried dimensions that correlate with other queried dimensions (Love et al., 2004). In Experiment 2, the roles the frequency of inference query play in learning and retention are examined. Previous eye-tracking studies have shown that learners in inference learning attend to previously queried dimensions even on trials in which another dimension is being queried (Rehder, Colner, & Hoffman, 2009). Then, a few inference queries may result in the same outcome as many queries. The current results will be useful in advancing category learning theories and guiding educational practice.

Experiment 1

In Experiment 1, inference and classification learning were compared with fifth-graders and undergraduates. Fifth-graders were studied as they learn about various categories around this grade, and they can follow the training procedures. Undergraduates were also examined to test whether the results generalize and to reinforce the link with the laboratory studies. Past work on children's category use (Hayes & Younger, 2004; Ross, Gelman, & Rosengren, 2005) and fits of a category learning model to developmental data (Gureckis & Love, 2004) suggest no strong differences between fifth-graders and undergraduates for the current tasks.

Table 1 summarizes the experimental phases and our main interests in Experiment 1.

Participants were trained on two contrasting categories consisting of items listed under Study item in Table 2 (A1-5 vs. B1-5). Each study item had five perceptual dimensions. For instance, if the first dimension is size with value 1 indicating small and value 2 indicating large, then B4 and B5 are both large. B4 and B5 match on the first, second, and fourth perceptual dimensions, as well as on the category label. The category labels were the names of the real sharks (e.g., Sixgill and Tiger sharks) rather than A and B. We used shark categories as fifth-graders were learning about sea animals when Experiment 1 was conducted. The five dimensions were mapped randomly onto the five binary-valued dimensions of

Table 1
Overview of Experiment 1

Phase	Main interest
Familiarization (15 trials)	
Training (60 trials)	Ease of learning - Overall accuracies Item type - Sensitivity to diagnosticity
Interruption (2 minutes)	
Test (20 trials)	Dimension - Acquisition of information with varying diagnosticity - Inference learners' acquisition of queried and non-queried information
Typicality (20 trials)	Item type - Treatment of transfer and study items - Sensitivity to diagnosticity
Retention (20 trials)	Dimension - Retention of information with varying diagnosticity - Inference learners' retention of queried and non-queried information

the sharks: habitat (near the surface or bottom), diet (fish or shrimp), litter size (a few or many pups), body size (small or large), and body shade (light or dark). The dimension values were assigned according to the real properties of the sharks used (described in *Materials*). Figure 2 highlights the five dimensions by displaying two sharks side by side with an opposite value on each dimension.

One unique aspect of Experiment 1 was that the categories combined family resemblance and rule-plus-exception structures. Whereas category A members tended to display value 1 across the five dimensions, category B members tended to display value 2 across the five dimensions. That is, members of the same categories displayed correlated properties. For example, whereas Sixgill sharks tend to be small and give birth to many pups, Tiger sharks tend to be large and give birth to only a few pups. At the same time, the first dimension was most diagnostic as it correctly classified 8 of 10 category members, compared to each of the other dimensions, which correctly classified 6 of 10 members. Using this category structure, learners' knowledge for high- and low-diagnostic dimensions can be compared. A number of converging measures were used to assess how attention was deployed and information was retained in inference and classification learning, and how direct queries affected inference learning.

Table 2
The Abstract Structure of the Items Used in the Training and Typicality Phases of Experiment 1

Study item	Dimension value	Transfer item	Dimension value
A1	A21111	T1	A21112
A2	A12112	T2	A12111
A3	A12211	T3	A11112
A4	A11221	T4	A12121
A5	A11122	T5	A11111
B1	B12222	T6	B12221
B2	B21221	T7	B21222
B3	B21122	T8	B22221
B4	B22112	T9	B21212
B5	B22211	T10	B22222

Note. During the training phase, participants were exposed to two contrasting categories (A1-A5 vs. B1-B5). The typicality phase included transfer items (T1-T10) in addition to study items. Each study item consisted of the category label denoted by A or B, and five perceptual stimulus dimensions with each value denoted by 1 or 2. For instance, if the first perceptual dimension is size with value 1 indicating small and value 2 indicating large, then both B4 and B5 are large. These two stimuli also match on the category label, and the second and fourth perceptual dimensions. T5 and T10 are the modal prototype of category A and category B, respectively.

Predictions

We predict that learners in classification learning will shift their attention to the most diagnostic regularity on the first dimension, and this information will become central to their category representations. In contrast, learners will focus on the prototypical nature of items in inference learning. We test these predictions by examining how learners treat different types of items. Study items A1 and B1 in Table 2 violate the most diagnostic regularity on the first dimension, but they have four category-typical values and share the most properties with the modal prototypes of their categories (i.e., A11111 and B22222). In contrast, study items A2-5 and B2-5 display the category-typical values on the first dimension and two of the remaining four perceptual dimensions. We specifically predict that learners in classification learning will make more training errors on A1/B1, which violate the regularity on the first dimension, than A2-5/B2-5, which follow the regularity on this dimension (Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004), and treat the regularity-violating items separately as exceptions (cf. Heit, 1998) even though these items are most similar to the category prototypes. Overall inference learning should result in fewer errors than classification learning.

The difference in learners' category representations under classification and inference was further examined by asking participants to rate how good an example each item in Table 2 was of its category after they learned about the sharks. We predict that A1/B1, which violate the regularity on the first dimension, should be relatively poor category examples for learners after classification training. In contrast, A1/B1 should be good category examples for learners after inference training as these items are the most typical study items. As shown in Table 2, the transfer items contained the category prototypes and other items that contained

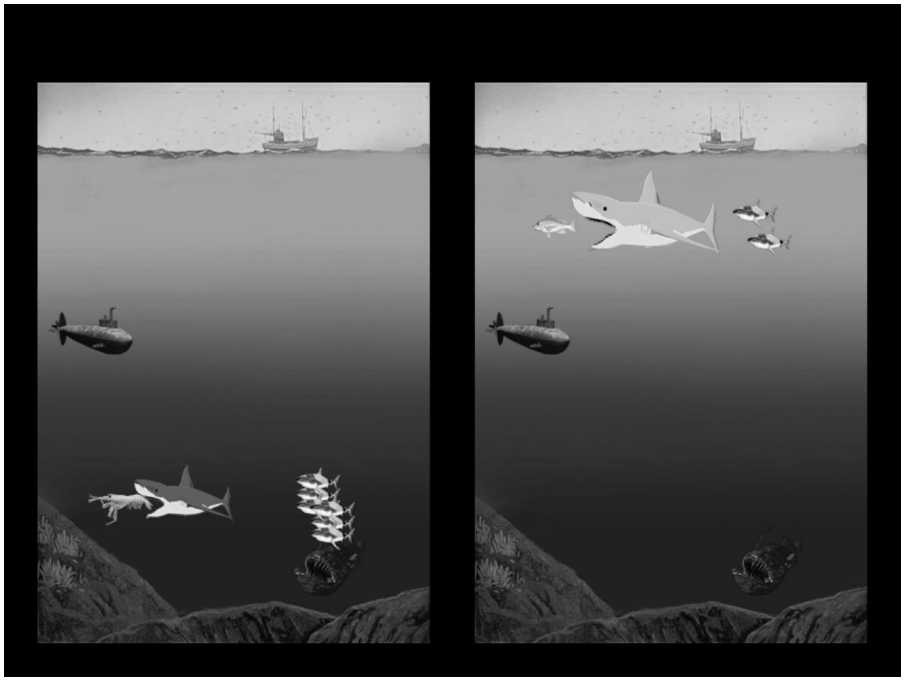

Figure 2. The prototypes (items T5 and T10 in Table 2) for the Sixgill (left) and Tiger (right) shark are shown. These two stimuli illustrate the contrasting values for the five perceptual dimensions: body size, body shade, diet, habitat, and the number of offspring.

more category-typical values than the study items. If learners attend to the property co-occurrences in each category during inference learning, they should rate the transfer items as overall better examples of the categories than the study items. Learners' typicality ratings after classification should be guided by the most diagnostic information on the first dimension.

Learners' category representations were also measured by their memory performances. A few minutes and multiple days after training, the learners' knowledge about the sharks' five dimensions was assessed. If learners in classification learning attend exclusively to the most diagnostic property, they will acquire little information about the low-diagnostic dimensions. In Experiment 1, the values of only three of five dimensions were queried during inference training. If learners merely memorize the correct value of each queried dimension in inference training (Johansen & Kruschke, 2005), the learners will not acquire information about the nonqueried dimensions. In contrast, if inference learning promotes the acquisition of property co-occurrences (Love et al., 2004), information about the nonqueried dimensions will be acquired.

Previous work has shown that dimensions not queried during inference training resulted in moderate knowledge, but the level of this knowledge did not differ (68% vs. 69% test accuracy) from that in classification learning (Anderson, Ross, & Chin-Parker, 2002). However, unlike our Experiment 1, all dimensions in their experiment were equally diagnostic, and they did not examine whether learners focused on a single dimension in classification learning. Moreover, they did not examine the learners' knowledge after multiple-day delay. We predict that whereas only information about the high-diagnostic dimension should be mastered following classification training, information about all dimensions, irrespective of being queried, should be mastered following inference training. We further predict that learners' memory for nonqueried dimensions following inference learning will be better than learners' memory for low-diagnostic dimensions after classification learning. To sum, we predict that inference training leads to more errors, and acquisition and retention of more category properties than classification training does.

Methods

Participants

Participants were 28 fifth-graders from St. Francis School of Austin and 54 University of Texas undergraduates. For the fifth-graders, there were 16 females and the mean age was approximately 10 years old. The general population of the undergraduates was approximately 68% female and the mean age was approximately 19 years old.

Apparatus

Each fifth-grader completed the experiment at St. Francis School of Austin using an Apple PowerBook computer operating in Mac OS X with a 15.2-inch TFT display. Each undergraduate was tested on a Pentium III computer operating in Windows 95 with a 15-inch CRT color display at the University of Texas at Austin. The resolution was 800 by 600 pixels.

Materials

The stimuli were computer animations of sharks swimming in the ocean. One animation cycle consisted of the shark appearing on

the right side of the display, swimming to the left side, and disappearing when it reached the left edge. The five binary-valued dimensions of the sharks and how they were mapped onto the category structure shown in Table 2 are explained in the introduction of Experiment 1.

One set of categories contrasted Sixgill and Tiger sharks. Relative to the Tiger sharks, the Sixgill sharks are common in deep water (90 to 600 m vs. surface to 340 m), often feed on shrimp (bottom dwelling invertebrate and bony fish vs. fish and almost anything), give birth to many pups (between 22 and 108 vs. from 10 to 80), are small (1.5 to 5 m vs. 3 to 6 m), and have dark body shade (dark gray or brown vs. grayish above and white below). Participants were informed that the sharks vary in their properties and thus the two categories' members could display overlapping properties. When Category A is the Sixgill shark, value 1 on each dimension in Table 2 signifies the value common to the Sixgill sharks. In this case, item T5 is a typical Sixgill shark that displays the category-typical values on all five dimensions (i.e., lives near the bottom, eats shrimp, delivers many pups, is small, and has dark shade) as shown in the left side of Figure 2, and item T10 is a typical Tiger shark displaying the category-typical values on all dimensions as shown in the right side of Figure 2.

The other set of categories contrasted Greenland and Soupfin sharks. The Greenland sharks are more common in deep water, eat fish more often, give birth to fewer pups, are larger, and have darker body shade than the Soupfin sharks. Information about the sharks was gathered from Enchanted Learning (2005) and Florida Museum of Natural History (2005).

Design and Procedure

The participants were randomly assigned to either the classification or inference condition. Seven to 33 days after completing the initial session, fifth-graders learned about a different set of sharks through a different learning mode. For example, fifth-graders who learned about the Sixgill and Tiger sharks through classification in the initial session learned about the Greenland and Soupfin sharks through inference in the second session. The delay had a high variance because we did not have a full control of the fifth-graders' and their teachers' schedules.

The experiment took about 20 minutes to complete in the initial session, which consisted of familiarization, training, interruption, test, and typicality phases, and a few minutes longer in the second session due to the additional retention phase after the same five phases as the initial session. Table 1 summarizes all of the phases. After each session, the fifth-graders were shown pictures of the actual sharks they had learned about. At the end of the second session, the fifth-graders were provided with a debriefing form, which included the five typical properties of each of the four sharks they learned. Each fifth-grader received a booklet of shark stickers at the completion of the initial session and a shark pen at the end of the second session. The undergraduates completed only the initial session and received course credit instead of shark paraphernalia. The instructions were displayed on the monitor at the start of each phase.

Familiarization. Prior to learning about the sharks, participants completed 15 familiarization trials, in which they were familiarized with the five stimulus dimensions. On each familiarization trial, the participants saw a pair of sharks side by side that differed on one of the five dimensions, and discriminated between the two possible values [e.g., "Which shark is larger? Left (Q) or right (P)?"]. The animation continued until the participants pressed the P or Q key to indicate the right or left side is correct, respectively. Then, a blank screen was displayed for 1000 ms, and the side that correctly depicted the shark reappeared for one animation cycle, together with the visual corrective feedback (e.g., "Right! The correct answer is P." or "Wrong! The correct answer is Q."). The participants also received auditory corrective feedback, a low-pitch tone for errors and a high-pitch tone for correct responses. Then, a blank screen was displayed for 1000 ms, and the next trial began. Each of the five dimensions was tested three times in a random order.

Training. Following familiarization, participants completed 60 training trials, in which they learned about the shark categories. Learning mode (classification or inference) was manipulated within participants for fifth-graders and between participants for undergraduates. Fifth-graders who were in the classification condition in the initial session completed the inference condition in the second session, and vice versa. Our main interests in the training phase were ease of learning under classification and inference training measured by overall training accuracies, and learners' sensitivity to diagnostic and prototypical information measured by the training accuracies on different item type (A1/B1 vs. A2-5/B2-5).

On each training trial, participants were shown two shark animations side by side as displayed in Figure 1A for classification learning and Figure 1B for inference learning. Whereas one side correctly described the shark, the other side did not. The correct and foil descriptions were randomly assigned to the left or right side. In classification learning, the foil and correct descriptions differed only on the text indicating the category membership (see Figure 1A): participants guessed the category label (e.g., Is this Tiger or Sixgill shark?). The reason for presenting two identical images at a time in classification learning was to equate the classification and inference learning trials as closely as possible. Typically, only one image appears with two or more possible category labels in a classification learning trial. In inference learning, the two descriptions differed only on the value of a queried dimension (see Figure 1B): participants guessed the dimension value (e.g., Is Tiger shark small or large?). Participants always guessed the value of one of the middle three dimensions during inference training.

The training phase was broken down into six blocks. Each training block consisted of a sequential presentation of the 10 study items under Study item in Table 2 in a random order. During each block of inference training, one of the middle three dimensions was queried for all 10 training items with the following constraints. The middle three dimensions were queried the same number of times. As in the majority of previous work (e.g., Chin-Parker & Ross, 2004; Yamauchi & Markman, 1998), the correct answer of the queried dimension was always typical of the shark's category (e.g., 1 for Sixgill and 2 for Tiger when Category A is Sixgill shark). For example, there was no inference learning trial for the first dimension of A1 in Table 2 (A?1111 → A21111, where ? signifies the inferred value), in which the correct answer would be inconsistent with the category-typical value. Such an inference trial is analogous to a classification trial of the prototype item belonging to the opposite category (?11111 → B11111). The

procedure in the training phase was identical to that in the familiarization phase.

Interruption. To prevent rehearsal of information from the training phase, a movie of 12 sharks swimming in the ocean sequentially was shown to the participants after training. Pictures of the Black-tip, Galapagos, Hammer Head, Horn, Lemon, Sandbar, Sharp Nose, Short-fin Maco, Whale, White, White-tip, and Zebra sharks were presented in a random order. Each shark was animated for 10,000 ms, with its name displayed at the bottom of the monitor.

Test. Participants completed 20 test trials after interruption. The main interests in the test phase were learners' acquisition of low- and high-diagnostic information in classification and inference, and learners' acquisition of queried and nonqueried information in inference. Participants' knowledge about the properties of the two categories from training was measured.

The test phase consisted of a sequential presentation of 20 text questions in a random order. Ten forced-choice questions asked the five typical properties of each of the two categories. For example, the text "Tiger sharks:" was presented above the two choices "A: tend to be smaller" and "B: tend to be larger" when the size of the Tiger shark was questioned. Another set of 10 questions asked the category typically associated with each of the 10 properties. For instance, the text "tend to be larger:" was displayed above the choices "A: Tiger sharks" and "B: Sixgill sharks" when the shark associated with the larger size was questioned. The correct choice was randomly assigned to the top (A) or bottom (B) position on each trial. To prevent learning during the test phase, no corrective feedback was provided. After participants responded, the text "Thank you" appeared beneath the choices for 2000 ms together with a brief high-pitch tone. Then, a blank screen was displayed for 2000 ms, and the next trial began.

Typicality. After the test phase, participants completed 20 typicality trials, in which they rated how good an example each shark in Table 2 was of its category. Participants were told to imagine a particular shark, such as a Tiger shark, and that a good example of a Tiger shark would look like the shark they just imagined. All 20 items in Table 2 were presented in a random order. The transfer items were placed in the categories according to the family resemblance structure. Typicality ratings measured the participants' knowledge organization.

On each typicality trial, an animated shark was presented with a rating scale. The text "Is this a good or a poor example of Sixgill sharks?" appeared above the scale when the presented shark was a member of the Sixgill shark. The ends of the scale were labeled "VERY GOOD" and "VERY POOR." To minimize artifacts, the polarities of the scale were counterbalanced such that for some participants the right end of the scale indicated a typical stimulus, but for others the left end did. The procedure in the typicality phase was identical to that in the test phase.

Retention. In the second session, the fifth-graders completed a retention phase after the five phases described so far. The retention phase was the repeat of the test phase from the initial session. The main interest in the retention phase was how well fifth-graders retained information about the shark categories learned through classification or inference in the initial session.

Results

All participants were included in the analyses. Our main interests were the fifth-graders' results from the training, test, typicality, and retention phases, shown in Figure 3. After presenting the fifth-graders' results, we present the undergraduates' results, shown in Figure 4. The two age groups showed the same general patterns of performances.

Fifth-Graders

No significant effects involving Session (initial or second) were found in the analyses of the fifth-graders' performances. Whether classification or inference learning was experienced first did not affect performances in Experiment 1. Thus, means were combined over this variable.

Training. A Learning Mode (classification or inference) by Item Type (A1/B1 or A2-5/B2-5) analysis of variance (ANOVA) was performed on the fifth-graders' training accuracies. As predicted in the *Predictions* section, the fifth-graders had higher accuracy under inference than classification training (.88 vs. .49), $F(1, 27) = 192.61$, mean standard error (*MSE*) = .02, $p < .01$. The effect size, measured by partial η^2, was .88, indicating that Learning Mode accounted for 88% of the effect plus error variance. As expected, the fifth-graders had higher training accuracy on rule-following but less prototypical A2-5/B2-5 than rule-violating but more prototypical A1/B1 (.72 vs. .65), $F(1, 27) = 8.47$, *MSE* = .02, $p < .01$, partial $\eta^2 = .24$. As predicted, there was a significant Learning Mode by Item Type interaction, $F(1, 27) = 4.73$, *MSE* = .02, $p = .04$, partial $\eta^2 = .15$. Figure 3A shows that, as predicted in the *Predictions* section, the fifth-graders' accuracy on rule-following A2-5/B2-5 was significantly higher than that on rule-violating A1/B1 in classification training, $t(27) = 2.82$, $p < .01$, $d = 0.81$ with 0.2, 0.5, and 0.8 indicating small, medium, and large effect sizes, respectively (Cohen, 1988). In contrast, the fifth-graders' accuracies on A1/B1 and A2-5/B2-5 did not differ significantly in inference training, $t(27) = .85$, $p = .40$.

Although the fifth-graders in classification had low overall training accuracies, their mean accuracy for rule-following A2-5/B2-5 (.59, *SD* = .14) in the last three training blocks was significantly above chance, $t(27) = 3.31$, $p = .01$, $d = 0.81$. Their mean accuracy for rule-violating A1/B1 (.38, *SD* = .23) was significantly below chance, $t(27) = -2.88$, $p < .01$, $d = 0.72$. The fifth-graders likely kept applying the learned rule to exception items A1/B1, which resulted in errors and discouraged them to rely exclusively on this rule.

Test. A response was "correct" when participants selected the property typical of the given category, or the category typically associated with the given property. No significant effects were found involving whether the participants predicted properties given categories, or predicted categories given properties. Thus, means were combined over this variable for further analyses.

Consistent with our prediction that inference training would lead to the acquisition of more category properties than classification training, the fifth-graders had higher test accuracy after inference than classification training (.84 vs. .62), $t(27) = 5.53$, $p < .01$, $d = 1.25$. To examine the fifth-graders' knowledge on different types of dimensions, we grouped the five dimensions into D1 (first dimension), D2-4 (middle three dimensions), and D5 (fifth dimen-

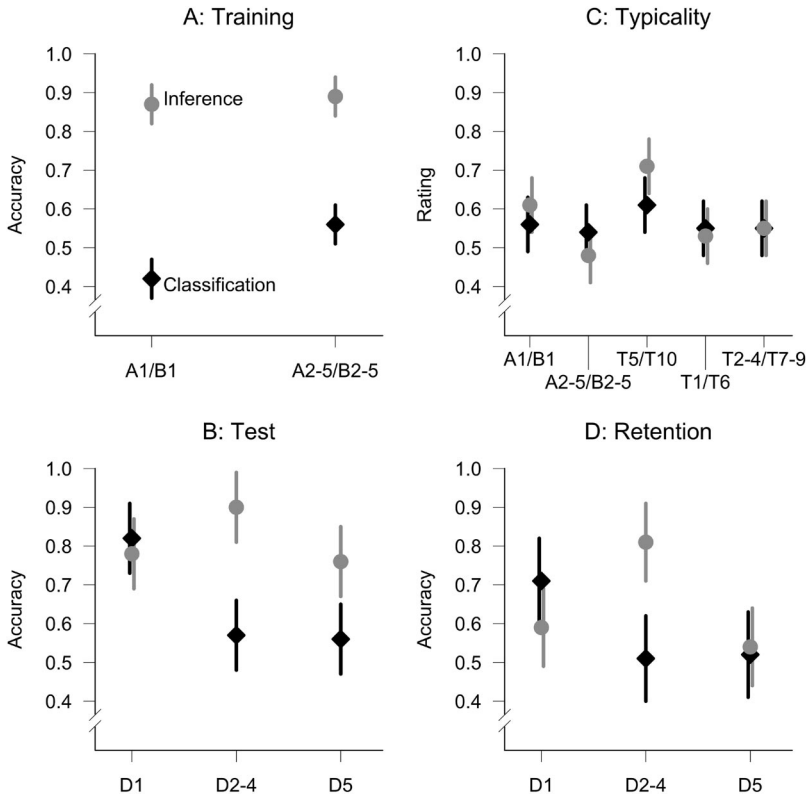

Figure 3. Fifth-graders' performances from Experiment 1 are shown. Error bars represent the 95% confidence intervals (see Loftus & Masson, 1994). [A] The inference procedure resulted in higher training accuracies than classification procedure. Items violating the rule on the first dimension (A1/B1) resulted in lower accuracy than items following the rule on the first dimension (A2-5/B2-5) under classification but not under inference training. [B] Only the most diagnostic dimension D1 resulted in an above chance test accuracy under the classification procedure. Queried dimensions D2-4 as well as nonqueried D1 and D5 resulted in above chance test accuracies under inference procedure. [C] In inference learning, studied items A1/B1 with four category-typical values and transfer prototype items T5/T10 with five category-typical values resulted in higher typicality ratings than studied items A2-5/B2-5 with three category-typical values, transfer items T1/T6 with three category-typical values, and transfer items T2-4/T7-9 with 3.67 category-typical values on average. The typicality ratings were flat in classification learning. [D] Only the most diagnostic dimension D1 in classification learning and the queried dimensions D2-4 in inference learning were retained after long-term delay.

sion). D1 was most diagnostic, and D2-4 were queried during inference learning. A Learning Mode by Dimension (D1, D2-4, or D5) ANOVA on the fifth-graders' test accuracies revealed that even with combined D2-4, the fifth-graders had higher test accuracy after inference than classification training (.81 vs. .65), $F(1, 27) = 9.89$, $MSE = .11$, $p < .01$, partial $\eta^2 = .27$. As expected, the fifth-graders' test accuracies on D1 (.80), D2-4 (.73), and D5 (.66) differed significantly from one another, $F(2, 54) = 3.42$, $MSE = .08$, $p < .05$, partial $\eta^2 = .11$. As predicted, there was a significant Learning Mode by Dimension interaction, $F(2, 54) = 4.83$, MSE = .10, $p = .01$, partial $\eta^2 = .15$. As predicted in the *Predictions* section, Figure 3B reveals above chance test performance for only D1 after classification learning, compared to strong performance for all dimensions after inference learning.

We further tested whether learners in inference learning garnered more information about low-diagnostic (nonqueried) dimensions than learners in classification learning (cf. Anderson et al., 2002) by comparing learners' performance on D5 (.76) after inference training to learners' performance on the second to fifth dimensions (.57) after classification training. As predicted, there was greater incidental learning under inference training, $t(27) = 2.20$, $p < .05$, $d = 0.63$. As can be seen in Figure 3B, the fifth-graders acquired information about nonqueried D1 equally well in inference and classification training ($t < 1$).

Typicality. Typicality ratings were mapped onto a 0 (atypical) to 1 (typical) scale with .5 as the midpoint. We first examined the typicality rating data using a regression analysis with three predictors: the number of category-typical values on D1 (0 or 1), D2-4 (1, 2, or 3), and D5 (0 or 1). If learners represent the categories around the most diagnostic information in classification learning, whether D1 has a category-typical value or not should be a good predictor of the fifth-graders' typicality ratings after classification learning. Inconsistent with this prediction, the three weights (D1 = .035, D2-4 = .034, D5 = .049) were all nonsignificant in fitting their ratings, $t < 1$, $t(27) = 1.38$, $p = .18$, and $t(27) = 1.41$, $p = .17$, respectively. Further, there were no differ-

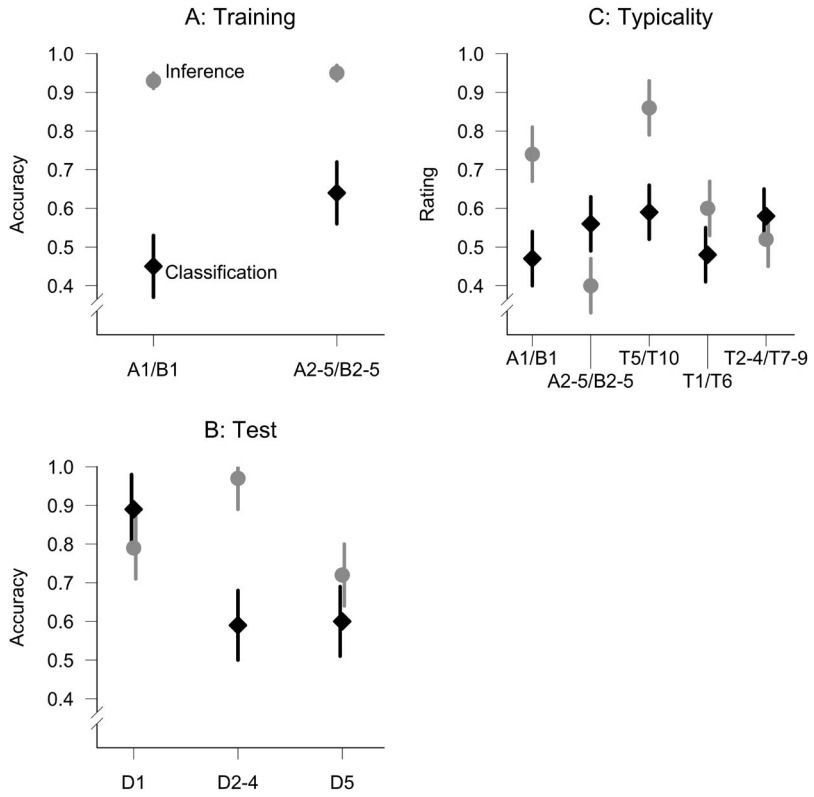

Figure 4. Undergraduates' performances from Experiment 1 are shown. Error bars represent the 95% confidence intervals (see Loftus & Masson, 1994). [A]: The inference procedure resulted in higher training accuracies than classification procedure. Items violating the rule on the first dimension (A1/B1) resulted in lower accuracy than items following the rule on the first dimension (A2-5/B2-5) under classification but not under inference training. [B]: The most diagnostic dimension D1 resulted in the highest test accuracy under classification procedure. Queried dimensions D2-4 resulted in the highest test accuracies under inference procedure. [C]: In inference learning, studied items A1/B1 with 4 category-typical values and transfer prototype items T5/T10 with five category-typical values resulted in higher typicality ratings than studied items A2-5/B2-5 with three category-typical values, transfer items T1/T6 with three category-typical values, and transfer items T2-4/T7-9 with 3.67 category-typical values on average. In classification learning, studied items A1/B1 and transfer items T1/T6, which violated the rule on the first dimension, resulted in lower typicality ratings than the other rule-following items (A2-5/B2-5, T2-4/T7-9, and T5/T10).

ences in the weights among D1, D2-4, and D5, $t < 1$ for all comparisons. Although the test phase results showed that the fifth-graders learned the first dimension in classification training, they did not rely on this dimension in typicality rating. As indicated by their low overall classification training performance, they had trouble using the rule-plus-exception strategy.

If learners acquire prototype information in inference learning, D1, D2-4, and D5 should all be good predictors of the fifth-graders' typicality ratings under inference learning. Partly consistent with this prediction, the weight on nonqueried but high-diagnostic D1 (.11) and the weight on queried D2-4 (.136) were significant in fitting their ratings, $t(27) = 5.35$, $p < .01$, and $t(27) = 2.33$, $p < .05$, respectively. However, the weight on nonqueried, low-diagnostic D5 (.055) was not significant, $t(27) = 1.37$, $p = .18$. The weight on D1 did not differ significantly from the weights on D2-4 and D5, $t < 1$ for both comparisons. The difference in weights between D2-4 and D5 approached significance, $t(27) = 1.85$, $p = .07$.

We further analyzed the fifth-graders' typicality ratings on different types of items. Item Type in the analysis of the typicality rating data included (A) training items A1/B1, which violated the rule on the first dimension but had four category-typical values, (B) training items A2-5/B2-5, which followed the rule on the first dimension and had three category-typical values, (C) transfer prototype items T5/T10 with five category-typical values, (D) transfer items T1/T6, which violated the rule on the first dimension but had three category-typical values on the middle three dimensions (akin to A1/B1), and (E) transfer items T2-4/T7-9, which followed the rule on the first dimension and had on average 3.67 category-typical values (akin to A2-5/B2-5).

Consistent with the results of the regression analysis, the fifth-graders did not show a systematic pattern in their typicality ratings following classification training (see Figure 3C). A one-way ANOVA on their typicality ratings with Item Type as a variable resulted in no significant main effect, $F(4, 108) = 0.64$, $MSE = 0.03$, $p = .64$. In contrast, there was a significant main effect of

Item Type for the fifth-graders' ratings after inference training (see Figure 3C), $F(4, 108) = 5.11$, $MSE = 0.04$, $p < .001$, partial $\eta^2 = .16$. As predicted if learners acquire prototype information in inference learning, the fifth-graders following inference training rated A1/B1 with 4 category-typical values as better category examples than A2-5/B2-5 with three category-typical values, $t(27) = 2.34$, $p < .05$, $d = 0.61$. They also rated never-before-seen prototype items T5/T10 as better category examples than items A2-5/B2-5 and T2-4/T7-9, $t(27) = 4.24$, $p < .001$, $d = 1.01$ and $t(27) = 3.39$, $p < .01$, $d = 0.78$, respectively. Further, they rated unstudied transfer items T2-4/T7-9 with 3.67 average category-typical values as better examples than studied items A2-5/B2-5 with three category-typical values, $t(27) = 2.27$, $p < .05$, $d = 0.47$. The fifth-graders in inference organized their category knowledge around prototype information.

Retention. The retention phase was the repeat of the training phase after delay. For the analyses of retention data, the fifth-graders were grouped according to the learning mode in the initial session. The distributions of delay were similar for the fifth-graders who learned through classification in the initial session ($M = 21$ days, standard error [SE] = 2 days, median = 18 days) and those who learned through inference in the initial session ($M = 20$ days, $SE = 2$ days, median = 18 days).

One of our main predictions was that learners would retain more category property information following inference than classification learning. As predicted, the fifth-graders had higher retention phase accuracy after inference than classification training (.71 vs. .55), $t(26) = 2.69$, $p < .05$, $d = 0.92$. The complete retention data pattern is shown in Figure 3D. As predicted, there was a significant Learning Mode by Dimension interaction, $F(2, 52) = 4.54$, $MSE = .07$, $p = .02$, partial $\eta^2 = .15$. Following inference training, although the fifth-graders retained knowledge of queried D2-4, they lost knowledge of nonqueried D1 and D5. Following classification training, the fifth-graders retained knowledge of high-diagnostic D1.

Undergraduates

The undergraduates' results mirrored the fifth-graders' except that the undergraduates showed more uneven typicality ratings than the fifth-graders did (see Figures 3 and 4).

Training. A Learning Mode by Item Type ANOVA was performed on the undergraduates' training accuracies. The undergraduates in the inference condition had higher training accuracies than those in the classification condition (.94 vs. .55), $F(1, 52) = 382.62$, $MSE = .01$, $p < .001$, partial $\eta^2 = .88$. As expected, the undergraduates had higher training accuracies on A2-5/B2-5 than A1/B1 (.79 vs. .69), $F(1, 52) = 13.75$, $MSE = .02$, $p < .001$, partial $\eta^2 = .21$. As predicted, there was a significant Learning Mode by Item Type interaction, $F(1, 52) = 9.30$, $MSE = .02$, $p < .01$, partial $\eta^2 = .15$. As shown in Figure 4A, whereas the undergraduates in the classification condition had significantly higher training accuracies on A2-5/B2-5 than A1/B1, $t(26) = 3.51$, $p < .01$, $d = 0.99$, those in the inference condition did not, $t(26) = 1.22$, $p = .23$. In the last three training blocks, whereas the mean accuracy for rule-following A2-5/B2-5 (.67, $SD = .21$) was significantly above chance, $t(26) = 4.19$, $p = .001$, $d = 0.99$, the mean accuracy for rule-violating A1/B1 (.44, $SD = .23$) was not significantly different from chance, $t(26) = -1.23$, $p = .22$. The undergraduates learned the rule but could not master the exceptions in classification training.

Test. Like the fifth-graders, the undergraduates had higher test accuracy after inference than classification training (.88 vs. .65), $t(52) = 5.18$, $p < .01$, $d = 1.16$. A Learning Mode by Dimension ANOVA was performed on the undergraduates' test accuracies. As expected, even with combined D2-4, the undergraduates had a higher test accuracy in the inference condition than in the classification condition (.83 vs. .69), $F(1, 52) = 5.70$, $MSE = .12$, $p < .05$, partial $\eta^2 = .10$. As expected, the undergraduates' test accuracies on D1 (.84), D2-4 (.78), and D5 (.66) differed significantly from one another, $F(2, 104) = 4.76$, $MSE = 0.09$, $p = .01$, partial $\eta^2 = .08$. As predicted, there was a significant Learning Mode by Dimension interaction, $F(2, 104) = 8.46$, $MSE = 0.09$, $p < .001$, partial $\eta^2 = .14$. Figure 4B reveals strongest test performance on D1 after classification learning and D2-4 after inference learning.

Typicality. Unlike the fifth-graders following classification training, the undergraduates in the classification condition relied on D1 the most in typicality ratings. The regression analysis revealed a significant weight on D1 (.116), $t(26) = 3.48$, $p < .01$, but not on D2-4 (.015) and D5 (.005), $t < 1$ for both, in fitting the undergraduates' typicality rating data in the classification condition. The weight on D1 was significantly larger than the weight on D2-4, $t(26) = 2.65$, $p < .05$, and marginally larger than the weight on D5, $t(26) = 1.97$, $p = .06$. The weights on D2-4 and D5 did not differ significantly, $t < 1$.

Undergraduates in the inference condition based their typicality decisions on the prototype information, more so than the fifth-graders under inference learning, as suggested by the significant weights on D1 (.089), D2-4 (.244), and D5 (.152), $t(26) = 2.23$, $p < .05$, $t(26) = 17.98$, $p < .001$, and $t(26) = 4.17$, $p < .001$, respectively. Queried D2-4 especially influenced the undergraduates' typicality ratings in the inference condition. The weight on D2-4 was significantly larger than the weights on D1 and D5, $t(26) = 4.03$, $p < .001$, and $t(26) = 2.95$, $p < .01$, respectively. The weights on D1 and D5 did not differ significantly, $t(26) = 1.39$, $p = .18$.

Mirroring the regression analysis, the analysis of the undergraduates' typicality ratings on different items in the classification condition showed that they organized the categories around the rule on the first dimension, and treated both A1/B1 and T1/T6 as deviant. Their typicality ratings on A1/B1, A2-5/B2-5, T5/T10, T1/T6, and T2-4/T7-9 differed significantly, $F(4, 104) = 3.13$, $MSE = 0.03$, $p < .05$, partial $\eta^2 = .11$. A planned comparison revealed that the undergraduates in the classification condition rated items A2-5/B2-5 and T2-4/T7-9, which followed the rule on the first dimension, as better category examples (.57 vs. .47) than items A1/B1 and T1/T6, which deviated the rule, $t(26) = 2.37$, $p < .05$, $d = 0.62$ (see Figure 4C).

For the undergraduates in the inference condition, prototype information was central to their category knowledge. Their typicality ratings of A1/B1, A2-5/B2-5, T5/T10, T1/T6, and T2-4/T7-9 differed significantly, $F(4, 104) = 28.21$, $MSE = 0.03$, $p < .001$, partial $\eta^2 = .52$. They rated more prototypical A1/B1 and T1/T6 as better category examples (.67 vs. .45) than less prototypical A2-5/B2-5 and T2-4/T7-9, $t(26) = 5.70$, $p < .001$, $d = 1.20$. This pattern of results is the opposite of the pattern observed for the undergraduates in the classification condition (see Figure 4C).

Further, the undergraduates in the inference condition rated prototypes T5/T10 as better category examples (.86 vs. .67) than A1/B1 and T1/T6 with category-atypical values on nonqueried dimensions, $t(26) = 3.32$, $p < .01$, $d = 0.96$, suggesting that they acquired information on the nonqueried dimensions, and this information played a role in their typicality ratings. Prototypes T5/T10 were also rated as better category examples (.86 vs. .45) than A2-5/B2-5 and T2-4/T7-9, $t(26) = 10.38$, $p < .001$, $d = 1.54$. Like the fifth-graders in inference learning, the undergraduates in the inference condition rated transfer items T2-4/T7-9 with more category typical values on average as better category examples (.52 vs. .40) than studied items A2-5/B2-5, $t(26) = 5.76$, $p < .001$, $d = 0.79$, also supporting their focus on prototype information.

Discussion

Experiment 1 showed that the inference task resulted in easier learning, and acquisition and retention of more category properties than the classification task. Participants in classification learning displayed little memory for dimensions other than the most diagnostic one (cf. Bott, Hoffman, & Murphy, 2007), suggesting that they focused exclusively on the most diagnostic dimension during training. Consistent with the prototype account (Love et al., 2004) but inconsistent with the rule account (Johansen & Kruschke, 2005), participants in inference training acquired information about both the queried and nonqueried properties, indicating that they did not focus exclusively on the queried dimensions during training (cf. Anderson et al., 2002). However, only queried dimensions were retained with delay.

Given that learners inferred multiple dimensions in inference learning but always predicted the category label in classification learning, one might propose that the "amount" of information given to the learners led to the performance differences in these two tasks. In other words, the learners in inference learning received a greater amount of direct information about category-property associations and thus performed better than those in classification learning. One way to equate the amount of information in the two tasks would be to always query the same dimension during inference training. In this situation, the current results suggest that the learners will acquire but not retain information about the nonqueried dimensions.

An alternative explanation might be that task difficulty, not procedure, resulted in the observed differences in category representations between classification and inference learning. As the training performance indicated, classification learning was more difficult than inference learning in Experiment 1. However, classification learning's focus on diagnostic information and inference learning's focus on prototypical information were found even when the two tasks were similar in difficulty (Chin-Parker & Ross, 2004). In the current work, we were interested in examining if the same categories were easier to learn through inference than classification learning, and thus did not equate the difficulty of the two tasks.

Although the learners' overall training accuracy in classification learning was low, they did learn and retain information about the diagnostic dimension. The undergraduates' classification accuracies for the rule-following items in Experiment 1 were comparable to undergraduates' classification accuracies in the previous studies using family resemblance structures (e.g., Anderson et al., 2002). The low overall classification performance is likely due to the presence of exception items. After discovering the rule, learners in classification make errors on the exception items by incorrectly applying the rule to these items. Learners also make errors on rule-following items when they falsely identify the rule-following items as exception items (Sakamoto & Love, 2006a). Many categories contain exceptions in the real world. The present results suggest that learning of such categories through classification procedure can hinder performance. The fifth-graders' typicality results suggest that students around this grade may especially struggle with learning categories containing exceptions through classification (Love & Gureckis, 2007), possibly due to the inability of their prefrontal cortex to support rule-plus-exception knowledge at this stage (Thompson-Schill, Ramscar, & Chrysikou, 2009).

There may be another benefit of inference procedure. Previous work has shown that ease of processing is a determinant of aesthetic pleasure (Reber, Schwarz, & Winkielman, 2004). If categories are more readily acquired through inference than classification learning, learners should prefer sharks acquired through inference training to those acquired through classification training. As predicted, of 11 fifth-graders who were asked "Which sharks did you like better, those from the last time or those from this time?" (six completed classification and five completed inference training in the initial session), 10 selected sharks learned through inference ($p = .01$, exact binomial, two-tailed). Ease of learning and liking of study materials associated with inference learning may play a critical role in students' motivation to learn (cf. Song & Schwarz, 2008).

To sum, inference training resulted in easier learning of more category information and better retention of learned materials than classification training. Only queried properties in inference learning and the most diagnostic property in classification learning were retained, indicating a role for explicit evaluation in establishing lasting memories. Further, fifth-graders liked materials learned through inference better than those learned through classification.

Experiment 2

Experiment 1 showed clear benefits of inference over classification learning. In Experiment 2, we examined whether the benefits from inference learning would require many queries or could be obtained with limited trials. In particular, we manipulated the frequency of query during inference learning and examined whether retention would improve when the dimension was queried more often, or querying over certain frequency would be enough to preserve memory.

The results from Experiment 2 will shed light on processes underlying inference learning and retention. If repeatedly answering the queries serves as rehearsal and leads to improved memory, then more queries may result in better retention. According to this account, learners retain information about queried dimensions because they respond to the queries. Alternatively, once learners are queried about a dimension, they may develop an expectation that they will be asked again about this dimension, and attend to this dimension even when they are not queried about this dimension. The results from eye tracking experiments of inference learning (Rehder et al., 2009) are consistent with the latter account. We predict that querying a few times should be enough to preserve

memory. Such results are unanticipated by the current theories of category learning (e.g., Johansen & Kruschke, 2005; Love et al., 2004).

Methods

Participants

Fifty University of Texas undergraduates who were not in Experiment 1 were recruited from the same population as that in Experiment 1.

Apparatus

Each participant used a 17-inch iMac. The resolution was 800 by 600 pixels.

Materials

The stimuli in Experiment 2 were animated sharks as in Experiment 1. The Sixgill and Tiger shark categories were used. The sharks in Experiment 2 were mapped onto the abstract category structure shown in Table 3. Unlike Experiment 1, each dimension in Experiment 2 was equally diagnostic in distinguishing members from different categories.

Design and Procedure

Each undergraduate completed two sessions. In the initial session, the undergraduates completed a familiarization, inference training, interruption, and test phases for course credit. Twelve to 33 days later, the participants completed a second session consisting of a retention phase, which was identical to the test phase in the initial session, for $7.

The design and procedure for each phase in Experiment 2 matched those in Experiment 1, except that there was only inference learning condition, and the frequency with which each of the five dimensions was queried varied. One dimension was queried 24 times, another dimension 18 times, another dimension 12 times, another dimension six times, and another dimension 0 time during training. The 60 training trials were broken down into three training blocks. The frequency of query for each dimension was distributed equally across the three training blocks.

Table 3
The Abstract Structure of the Items Used in the Training Phase of Experiment 2

Study item	Dimension value
A1	A11112
A2	A11121
A3	A11211
A4	A12111
A5	A21111
B1	B22221
B2	B22212
B3	B22122
B4	B21222
B5	B12222

Note. During the training phase, participants learned about two contrasting categories (A1-A5 vs. B1-B5) through inference procedure. Each study item consisted of the category label denoted by A or B, and five perceptual stimulus dimensions with each value denoted by 1 or 2. For instance, if the first perceptual dimension is size with value 1 indicating small and value 2 indicating large, then both A1 and A2 are small. These two stimuli also match on the category label, and the second and third perceptual dimensions.

Results

All participants were included in the analyses. Our main interests were the performances on each dimension in the training, test, and retention phases. Figure 5 summarizes the results.

Training

The differences in the training accuracies for dimensions queried 6, 12, 18, and 24 times did not reach significance, $F(3, 147) = 2.11$, $MSE = .03$, $p = .1$. The difference in the training accuracies for dimensions queried 24 times and six times was not significant, $t(49) = 1.63$, $p = .11$.

Test

Participants' test accuracies for dimensions queried 0, 6, 12, 18, and 24 times differed significantly, $F(4, 196) = 13.24$, $MSE = .09$, $p < .01$, partial $\eta^2 = .21$. When only the dimensions that were queried were compared, the differences in the test accuracies did not reach significance $F(3, 147) = 2.36$, $MSE = .07$, $p = .07$. Unlike Experiment 1, the participants did not learn about the nonqueried dimension in Experiment 2. Their test accuracy on the dimension queried 0 time did not differ significantly from the chance level of .5 ($p = .8$). Their test accuracy on each of the other dimensions was significantly above chance ($p < .01$ for each comparison).

Retention

Participants' retention accuracies for dimensions queried 0, 6, 12, 18, and 24 times did not differ significantly, $F(4, 196) = 1.89$, $MSE = .14$, $p = .11$. Mirroring test performance, whereas performance on queried dimensions was above chance ($p < .01$ for each comparison), performance on the nonqueried dimension did not differ significantly from chance ($p = .29$). As shown in Figure 5, for queried dimensions, the number of queries during training had surprisingly little effect on participants' retention accuracy.

Relationship Between Training and Later Performances

The advantage of inference over classification training in Experiment 1 suggests that conditions that result in fewer errors can result in better learning and memory. In Experiment 2, the training accuracy correlated positively with the test accuracy for each queried dimension ($r = .39$, $p < .01$ for the dimension queried 6 times; $r = .43$, $p < .01$ for 12; $r = .41$, $p < .01$ for 18; $r = .49$, $p < .01$ for 24). The training accuracy also correlated positively with the retention accuracy on each queried dimension except for the one queried least frequently ($r = .18$, $p = .22$ for the dimension queried 6 times; $r = .46$, $p < .01$ for 12; $r = .3$, $p < .05$ for 18; $r = .35$, $p = .01$ for 24).

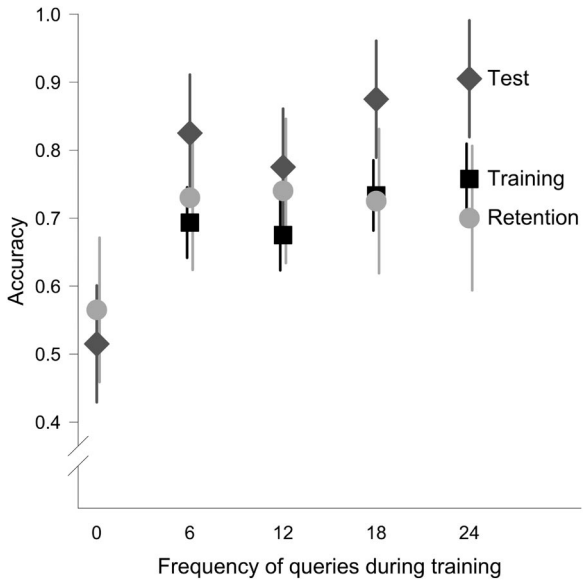

Figure 5. Undergraduates' performances from Experiment 2 are shown. There are no training data for the dimension queried 0 time during training. Error bars represent the 95% confidence intervals (see Loftus & Masson, 1994). The frequency of inference queries, once above zero, has surprisingly little influence on training, test, and retention accuracies.

Discussion

The frequency of inference query, once above zero, had little effect on learners' training, test, and retention performances. The lack of significant differences in training accuracies between dimensions queried six times and 24 times suggest that learning takes place quickly through inference procedure. The lack of large differences in test and retention performances between infrequently and frequently queried dimensions suggests that a query of a dimension serves as a signal to the learner to attend to the category-property relationships for that dimension even on trials involving other dimensions. Querying a few times is enough to induce repeated retrieval during training, which plays an important role in retention (Karpicke & Roediger, 2007).

To our surprise, participants did not acquire information about the nonqueried dimension in Experiment 2. The inference task in Experiment 2 was more demanding than that in Experiment 1 (.72 training accuracy in Experiment 2 vs. .94 for undergraduates in Experiment 1). Remembering information about the dimensions queried less frequently in Experiment 2 might have consumed additional cognitive resources, preventing the participants from attending to the nonqueried dimension. Further, whereas there was only one nonqueried dimension in Experiment 2, there were two nonqueried dimensions in Experiment 1.

Another finding from Experiment 2 was that making errors correlated with worse test and retention performances. This finding seems to contradict the classic theory that errors drive learning and changes in memory (Rescorla & Wagner, 1972). Perhaps participants experienced source confusion when there was conflict between response and feedback. For example, they might forget if they responded a Tiger shark but it was a Sixgill shark, or it was the other way around. Consequently, they might remember an incorrect stimulus-response mapping.

General Discussion

In the current study, we examined the nature of learning and retention in inference and classification tasks. In Experiment 1, we also extended the basic findings from inference and classification learning studies with undergraduates to fifth-graders in a classroom setting. In Experiment 2, we focused on the role queries play during inference learning in shaping category acquisition and retention. After presenting an overview of the current findings, we discuss the theoretical and educational implications of the current work.

Findings From the Current Work

Inference procedure led to easier learning of more category information, and better retention and liking of learned materials than classification procedure in Experiment 1. There was surprisingly little benefit of additional inference queries for retention in Experiment 2, indicating that only a limited number of queries are required to boost long-term retention. Next, we discuss our key findings in relation to the questions we have posed in the introduction.

Organization of category knowledge. The participants organized their category knowledge differently under inference and classification learning in Experiment 1. The fifth-graders, like the undergraduates, organized their knowledge around the prototypes in inference learning. The undergraduates in classification learning organized their knowledge around the most diagnostic information. Although the fifth-graders retained the most diagnostic information under classification learning, this information did not influence their typicality ratings.

Fifth-graders' results from Experiment 1 provided new insights into the nature of retention in inference and classification learning. Explicit evaluation played a key role in retention: only queried properties in inference learning and the diagnostic property in classification learning were retained with delay. The retention of the most diagnostic information after classification learning is analogous to the result in which queried dimensions were retained after inference learning if one assumes that learners actively engage in hypothesis testing involving the most diagnostic dimension during classification learning. Overall, learners retained more knowledge after inference than classification learning.

Ease of learning. Learning was easier through inference than classification procedure in Experiment 1. In Experiment 2, making more training errors correlated with worse test and retention performances, contradicting error-driven learning theories (e.g., Rescorla & Wagner, 1972).

Processes underlying classification and inference learning. The learners in classification focused exclusively on the most diagnostic information during training and retained no information about the other dimensions. The learners in inference acquired information about both the queried and nonqueried properties in Experiment 1 (cf. Anderson et al., 2002). However, the results from Experiment 2 suggested that the processing of nonqueried dimensions could be prevented when cognitive resources were limited.

In Experiment 2, only a few queries were enough for the retention of category-property relationships. The surprisingly little benefit of additional queries, especially when long-term retention

is considered, suggests that it is not the repeated answering to queries that leads to improved retention. Instead, when queried about a dimension, learners may develop an expectation that they will be asked again about this dimension and rehearse information about this dimension on every inference trial, resulting in little effect of query frequency on retention.

Implications for Theories of Category Learning

Existing theories of category learning do not specify how queries shape retention (cf. Sakamoto & Matsuka, 2007). These theories do not address how making errors can both facilitate and hinder memory. Our results can guide the further development of these theories.

The role of queries in learning and retention. Departing from memory models (e.g., Shiffrin & Steyvers, 1997), most category learning models assume that all dimension values are encoded on each trial with attentional mechanisms determining the relative weighting of stimulus dimensions (e.g., Kruschke, 1992; Love et al., 2004). Clearly, the attentional mechanisms in category learning theories need to be elaborated to address the role queries play in shaping attention. For example, models need to address our finding that people learn information about nonqueried dimensions, which is not emphasized at study, but capacity limitations may prevent learners from entertaining the nonqueried dimensions. Further, models need to consider our finding that information about the dimensions that are queried less frequently is remembered as well as information about the dimensions that are queried more frequently. Attentional processes may operate across multiple trials. For instance, once a learner is queried about a dimension (e.g., "Does a fish swim?"), the learner continues to attend to the dimension on subsequent trials that ask about other dimensions that are related (e.g., "Does a fish have fins?") when cognitive resources are available.

The role of errors in learning and retention. The finding that more errors can correlate with worse acquisition and retention of knowledge can also guide the refinement of category learning models. Many models of category learning assume that errors play a central role in learning (e.g., Rescorla & Wagner, 1972) by leading to greater attention to error-producing items (e.g., Mackintosh, 1975), and use error-minimization techniques to drive learning (e.g., Kruschke, 1992; Love et al., 2004). The current work however has shown that errors do not always lead to improved memory. These models of category learning need to address when errors help or hinder memory storage.

Implications for Education

Our results suggest that a few inference queries can lead to better retention of knowledge than many classification queries. Classroom teaching should emphasize reasoning from the category to multiple properties rather than from a set of properties to the category.

Errors during category learning. The advantages of inference over classification learning suggest that making errors are not always helpful in establishing memories. Perhaps making many errors during category learning can lead to source monitoring problems. For example, the learner might forget whether they responded a Tiger shark but it was a Sixgill shark, or it was the other way around. Such errors are often common in classification learning.

Testing and additional study. One technique for improving retention in education is overlearning, or additional practice of well-learned materials (Driskell, Willis, & Copper, 1992; but see Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005). However, some researchers (Nungester & Dchastel, 1982; Roediger & Karpicke, 2006) have proposed that testing on material can be more beneficial to establishing memories than additional practice. The benefit of testing parallels the advantage of inference learning. Directly testing properties in inference training can improve learning and retention.

Karpicke and Roediger (2007) have further proposed that although additional studying may not improve retention, additional testing will. The present results suggest that additional testing may not be necessary for long-term retention of learned materials if testing is designed to promote repeated processing of information even when the learner is not directly tested.

Direct instruction and discovery learning. For learning the category-property associations, the inference procedure is like direct instruction in that learners are asked directly about properties associated with the category. The classification procedure is like discovery learning in that learners discover the category-property associations on their own. The advantage of inference over classification may be related to the finding that direct instruction can lead to successful learning by many more third- and fourth-grade students than discovery learning (Klahr & Nigam, 2004). Inference procedure, like direct instruction, guides the learner on what needs to be acquired. The lack of guidance in classification and discovery learning may lead to many errors, which may outweigh the benefit of active processing in these tasks.

Unsupervised category learning. Unlike the supervised inference and classification training in the current work, classroom learning likely involves unsupervised learning as well, in which learners receive no corrective feedback. Although unsupervised and supervised learning have been compared to each other (e.g., Love, 2003), more work is needed to understand how learning progresses when unsupervised and supervised learning episodes are interwoven. Given that testing on material can improve memory in the absence of feedback (Roediger & Karpicke, 2006), simply querying properties in inference learning without corrective feedback might improve retention of category properties. Moreover, including unsupervised learning trials within supervised inference and classification learning may have differential effects on learning. Whereas unsupervised trials should always benefit inference learning by reinforcing the extracted prototype, unsupervised presentation of the items that display the amodal value on the diagnostic dimension (such as A1/B1 in Table 2) should retard classification learning.

Final Notes

Of course, inference learning is not always advantageous. Whether inference learning is more efficient than classification learning depends on the structure of the categories. Inference learning is advantageous when categories have family resemblance structure and learning involves discovering relationships among properties within categories as in the present experiment. In contrast, the prototype knowledge in inference learning can interfere

with learning about a nonlinear category structure, in which the category prototypes do not help in discriminating members of different categories (Yamauchi et al., 2002). However, nonlinear category structures are rare outside of the laboratory (Murphy, 2002).

Furthermore, whether the inference task aids learning depends on what information is queried during training. For example, the learners in Nilsson and Olsson's (2005) experiment showed no learning when the inference task included trials involving exception features that were more typical of the opposing category. These results suggest that it is important to constrain inference training to query only the category-typical values and to tailor the training procedure to match the structure of the categories.

Finally, individual differences in cognitive style may influence whether inference learning is beneficial or not. For instance, individuals with high cognitive ability may do equally well in classification and inference procedure. Future work should examine the potential importance of aptitude-treatment interactions.

With these caveats in mind, our messages for educators are:

- Classroom exercises should emphasize reasoning from the category to multiple properties (i.e., inference) rather than from a set of properties to the category (i.e., classification).
- Teachers should ask about important properties during inference training just a few times, but may NOT want to test properties that are not important to prevent cognitive overload.
- Testing should be designed to promote repeated processing of information even when the learner is not directly tested on the information.
- Learning procedure should not result in many errors to help memory storage, as well as to increase the learners' liking of study materials and possibly motivation to learn.

References

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition, 30,* 119–128.

Bott, L., Hoffman, A. B., & Murphy, G. L. (2007). Blocking in category learning. *Journal of Experimental Psychology: General, 136,* 685–699.

Callanan, M. (1985). How parents label objects for young children: The role of input in the acquisition of category hierarchies. *Child Development, 56,* 508–523.

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 216–226.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). New York: Academic Press.

Driskell, J. E., Willis, R. P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77,* 615–622.

Enchanted Learning. (2005). Retrieved from http://www.enchantedlearning.com/

Florida Museum of Natural History. (2005). Retrieved from http://www.flmnh.ufl.edu/fish/sharks/sharks.htm

Gureckis, T. M., & Love, B. C. (2004). Common mechanisms in infant and adult category learning. *Infancy, 5,* 173–198.

Hayes, B. K., & Younger, K. (2004). Category-use effect in children. *Child Development, 75,* 1719–1732.

Heit, E. (1998). Influences of prior knowledge on selective weighting of category members. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 712–731.

Johansen, M. K., & Kruschke, J. K. (2005). Category representation for classification and feature inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1433–1458.

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15,* 661–667.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490.

Love, B. C. (2003). The multifaceted nature of unsupervised category learning. *Psychonomic Bulletin & Review, 10,* 190–197.

Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science, 14,* 195–199.

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience, 7,* 90–108.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of human category learning. *Psychological Review, 111,* 309–332.

Lynch, E. B., Coley, J. D., & Medin, D. L. (2000). Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory & Cognition, 28,* 41–50.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimulus with reinforcement. *Psychological Review, 82,* 276–298.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129,* 592–613.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors and feedback. *Psychonomic Bulletin & Review, 14,* 225–229.

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

Nilsson, H., & Olsson, H. (2005). Categorization vs. inference: Shift in attention or in representation? In *Proceedings of the 27th Annual Conference of the Cognitive Science Society.* Stresa, Italy: Cognitive Science Society.

Ninio, A. (1980). Ostensive definition in vocabulary teaching. *Journal of Child Language, 7,* 565–573.

Ninio, A., & Bruner, J. S. (1978). The achievement and antecedents of labelling. *Journal of Child Language, 5,* 1–15.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,* 53–79.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74,* 18–22.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 548–568.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied, 14,* 101–117.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality & Social Psychology Review, 8,* 364–382.

Rehder, B., Colner, R. M., & Hoffman, A. B. (2009). Feature inference learning and eyetracking. *Journal of Memory and Language, 60,* 393–419.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing

memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210.

Rohrer, D., Taylor, K., Pashler, H., Wixted, J. T., & Cepeda, N. J. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology, 19,* 361–374.

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

Ross, B. H., Gelman, S. A., & Rosengren, K. S. (2005). Children's category-based inferences affect classification. *British Journal of Developmental Psychology, 23,* 1–24.

Rule, A. C. (2007). Mystery boxes: Helping children improve their reasoning. *Early Childhood Education Journal, 35,* 13–18.

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General, 133,* 534–553.

Sakamoto, Y., & Love, B. C. (2006a). Vancouver, Toronto, Montreal, Austin: Enhanced oddball memory through differentiation, not isolation. *Psychonomic Bulletin & Review, 13,* 474–479.

Sakamoto, Y., & Love, B. C. (2006b). Sizable sharks swim swiftly: Learning correlations through inference in a classroom setting. In R. Sun and N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Vancouver, Canada: Cognitive Science Society.

Sakamoto, Y., & Love, B. C. (2009). You only had to ask me once: Long-term retention requires direct queries during learning. In N. Taatgen, H. van Rijn, L. Schomaker, and J. Nerbonne (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Amsterdam, Netherlands: Cognitive Science Society.

Sakamoto, Y., & Matsuka, T. (2007). Incorporating forgetting in a category learning model. In *Proceedings of International Joint Conference on Neural Networks.* Orlando, FL: IEEE.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychonomic Bulletin & Review, 3,* 145–166.

Song, H., & Schwarz, N. (2008). If it's hard to read, it's hard to do: Processing fluency affects effort prediction and motivation. *Psychological Science, 19,* 986–988.

Thompson-Schill, S. L., Ramscar, M., & Chrysikou, E. G. (2009). Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Science, 18,* 259–263.

Wixted, J. T. (2005). A theory about why we forget what we once knew. *Current Directions in Psychological Science, 14,* 6–9.

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 585–593.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39,* 124–148.

Received March 1, 2009
Revision received July 29, 2010
Accepted September 10, 2010 ■