JOURNAL OF COGNITION AND DEVELOPMENT, *8*(1), 93–111

Looks Aren't Everything: 24-Month-Olds' Willingness to Accept Unexpected Labels

Vikram K. Jaswal
Department of Psychology
University of Virginia

Ellen M. Markman
Department of Psychology
Stanford University

A label can efficiently convey nonobvious information about category membership, but this information can sometimes conflict with one's own expectations. Two studies explored whether 24-month-olds ($N = 56$) would be willing to accept a category label indicating that an animal (Study 1) or artifact (Study 2) that looked like a member of one familiar category was actually a member of a different familiar category. Results showed that children were receptive to these unexpected labels and used them as the basis for inference. These findings indicate that linguistic information can lead even toddlers to "disbelieve their eyes."

The testimony of others represents a crucial source of information for children and adults alike (Coady, 1992; Harris, 2002; Reid, 1764/1997). It readily serves as a more efficient means of gaining knowledge than first-hand observation: Hearing a radio announcer explain that an event has been canceled can save us the trouble of going to the event and discovering this for ourselves. In some domains, testimony is not only a more efficient way of acquiring knowledge than first-hand observation, it may be the only way. When it comes to learning about events that took place before we were born, for example, we have to rely on what other people tell us (live or via a symbolic medium). Finally, testimony can sometimes lead us to revise a

Correspondence should be sent to Vikram Jaswal, Department of Psychology, 102 Gilmer Hall, P.O. Box 400400, University of Virginia, Charlottesville, VA 22904-4400. E-mail: jaswal@virginia.edu

belief or expectation that we generated from personal observation. Even though an eel looks like a snake, for example, most of us learn from testimony (rather than observation) that it is actually a fish. The studies presented in this article focus on this last point: To what extent are toddlers willing to "disbelieve their eyes" about the category in which an object belongs on the basis of what someone tells them?

One of the most important functions of categories is that they provide a rich knowledge base from which inferences can be drawn (Markman, 1989). For example, if you know that birds hatch from eggs and that a particular animal is a bird, then you can infer that it hatched from an egg even if you didn't witness its birth. Much of the research on categorization in children has focused on how they form categories on the basis of similarities they detect themselves (for a review, see Rakison & Oakes, 2003). For example, Quinn, Eimas, and Rosenkrantz (1993) showed that very young infants can form categories by detecting perceptual similarities between objects. In one study, 3.5-month-olds who were habituated to a series of pictures of cats dishabituated when shown a picture of a dog. Other research has shown that children can also form categories by detecting nonobvious similarities between objects, such as causal or functional similarities (e.g., Gopnik & Sobel, 2000; Kemler Nelson, Russel, Duke, & Jones, 2000). In Gopnik and Sobel's (2000) study, for example, 2.5-year-olds treated two perceptually dissimilar objects as members of the same category if both caused a machine to light up and play music.

However, in addition to learning about and forming categories on the basis of observation, children can also do so from testimony. Consider how they learn that penguins are birds. Penguins don't look much like typical birds, and most children don't spend enough time with penguins to be able to detect for themselves the reasons why penguins are birds. Nevertheless, between 12 and 24 months, children begin to recognize many atypical exemplars as members of familiar categories (Meints, Plunkett, & Harris, 1999) — presumably because they have heard people use familiar category labels when referring to them (Adams & Bullock, 1986).

Indeed, the labels that are applied to objects can have a profound influence on the categories that children form (e.g., Markman & Hutchinson, 1984; Nazzi & Gopnik, 2001; Waxman & Markow, 1995; Welder & Graham, 2001). In Waxman and Markow's (1995) study, for example, 12-month-olds heard a series of perceptually disparate exemplars from the same superordinate category (e.g., animals) referred to with a single label or without a label. When later shown an exemplar from that category and an exemplar from a different superordinate category (e.g., a vehicle), children in the label condition showed greater interest in the new superordinate category, whereas those in the no-label condition did not. Waxman and Markow argued that a shared label can serve as an invitation to form a category that children would not otherwise form.

Recently, Graham, Kilbreath, and Welder (2004) showed that a shared label can also lead infants to draw inferences between two objects they normally would not draw inferences between. In that study, 13-month-olds saw an experimenter demonstrate that a novel artifact had an interesting, nonobvious property that could be elicited by a particular action (e.g., squeezing an object would produce a squeak). When provided with another, highly perceptually similar object, these infants spontaneously performed the action, apparently attempting to elicit the nonobvious property. In contrast, when provided with an object that was low in perceptual similarity to the original object, infants did not (see also Baldwin, Markman, & Melartin, 1993). When the original object and the low-similarity object were both referred to by the same novel name, however, infants did attempt to perform the target action on both — as if they now considered them to be members of the same category and, therefore, to share the same nonobvious property.

In addition to leading children to form an entirely new category, a label can have an even more powerful effect: It can also lead them to reclassify an object from one known category to another known category (e.g., Gelman & Coley, 1990; Gelman, Collman, & Maccoby, 1986; Gelman & Markman, 1986, 1987; Sloutsky & Fisher, 2004). In Gelman and Markman (1986), for example, 4-year-olds were shown a picture of a squirrel and told, "This squirrel eats bugs." They were then shown a picture of a rabbit and told, "This rabbit eats grass." Finally, they were shown a picture of a squirrel that looked very much like a rabbit (e.g., it had very long ears) and were asked, "See this squirrel? Does it eat bugs, like this squirrel, or does it eat grass, like this rabbit?" Even though perceptual similarity is normally a very good cue to category membership (e.g., Jones & Smith, 1993; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), children tended to respond on the basis of the label, inferring that the rabbit-like squirrel would eat bugs like other squirrels.

In a later study, Gelman and Coley (1990) simplified the procedure for use with 2.5-year-olds. Children were shown a typical exemplar from one familiar category and were reminded that it possessed a familiar, nonobvious property (e.g., they were shown a dog and told that it barked). The experimenter then showed four test pictures, one at a time: One typical and one atypical member of the same category as the labeled category (e.g., a Golden Retriever, a Chihuahua), and one typical and one atypical member of a different category (e.g., a white lamb, a dog-like lamb). Each picture was referred to neutrally or was labeled with its appropriate label (e.g., "dog" or "lamb"), and children were asked whether it possessed the property in question (e.g., "Does this bark?" or "Does this dog/lamb bark?"). In the no-label condition, children relied on appearance to make the inference: For example, they inferred that the dog-like lamb barked. When the same pictures were labeled, however, children responded on the basis of the category label. For example, they agreed that the Golden Retriever and Chihuahua (both referred to as "dogs") barked, and they denied that either of the lambs (both referred to as "lambs") did so.

Children's willingness to make an inference on the basis of an object's category label, even when the label conflicts with its appearance, is noteworthy for a number of reasons. First, consistent with the results from Waxman and Markow (1995) and Graham et al. (2004) described earlier, it suggests that children expect that something deeper than surface similarity is responsible for category membership (e.g., Bloom, 2000; Gelman, 2003; Medin & Ortony, 1989). Exactly what makes a dog a dog, or a lamb a lamb, may be unknown (and unknowable; Gelman, 2003), but a category label can serve as a placeholder that a reason exists.

Second, results from Gelman and Coley's (1990) study suggest that even 2.5-year-olds recognize that appearances can be misleading: Just because something looks like a dog does not mean that it is a dog; it could be a lamb. Finally, their results suggest that children are deferential to the testimony of an adult on matters of categorization: Even though the labels conflicted with children's own observations, and even though the speaker did not provide any explanation for these disparities, 2.5-year-olds made label-based rather than perceptually based inferences.

In the studies reported here, we asked whether 24-month-olds would also be willing to give up a compelling, perceptually based classification in favor of an unexplained, label-based one. There is some reason to expect that they would not. Banigan and Mervis (1988) found that 24-month-olds who overextended a known word (e.g., used "ball" to refer to a candle) were unlikely to correct that overextension if the speaker simply provided the correct label (see also Mervis, 1984). In fact, in diary data from her son's first 2 years, Mervis (1987; Mervis, Pani, & Pani, 2003) found that he often ignored corrections from his parents, behaving as if "categories should not be formed on the basis of adult labeling patterns alone" (Mervis, 1987, p. 213).

Mervis's (Banigan & Mervis, 1988; Mervis, 1984, 1987; Mervis et al., 2003) interest was in whether toddlers who heard an adult use an unexpected label would later understand and produce that label. This required that children accept the unexpected label and then later remember why it applies and why their original classification does not. In contrast, our interest is in the first step only — namely, whether 24-month-olds are willing to accept that an unexpected classification applies in the first place. If they are, then they may have an implicit understanding that appearances can be misleading. (How else could something look like one thing but actually be something else?) In addition, if a label can lead them to reclassify an object, this suggests that they recognize that the label someone uses to refer to an object could be a better source of information about the category in which an object belongs than their own personal observations.

To analyze this question with 24-month-olds, we used an inductive inference paradigm adapted from Gelman and Markman (1986). Children watched as a researcher used small props to act out an activity associated with each of two familiar categories (e.g., a cat drank milk, a dog chewed on bones). They were then presented with a hybrid object designed to look more like a member of one category than the other. They were asked to show, by acting out with the props, which activity the hybrid engaged in (e.g., did it drink milk or eat bones?). Some children heard the hybrid labeled with a counterintuitive label (e.g., a cat-like animal was referred to as "this dog"), and some children heard it referred to without a label (e.g., the same animal was referred to as "this one"). Of interest was whether children who heard the counterintuitive label would make the hybrid perform the opposite activity from those who did not hear the label.

STUDY 1

Method

Participants. Thirty-two 24-month-olds (*M* = 24 months, 17 days; range = 22;29 to 26;9) participated in a single 10- to 15-min session. Sixteen were girls, and 16 were boys. Ten additional toddlers were tested, but their data were not included due to extreme fussiness resulting in an inability to complete at least half of the session (5), experimenter error (3), camera failure (1), or exclusive selections to the child's right side (1).

Design. Children were randomly assigned to a label or no-label condition, resulting in 16 participants per condition, balanced for gender. The average age was 24;18 in the label condition and 24;16 in the no-label condition.

Materials. Eight animals from familiar categories were grouped into four pairs based roughly on similar sizes and body shapes: cat-dog, horse-cow, pig-bear, and bird-fish. To confirm that these categories were familiar, prior to each session, parents completed a brief vocabulary checklist that asked the level of their child's understanding of each label: definitely understands, maybe understands, or does not understand. Over 60% of the children in each condition were reported to definitely understand all eight labels (11/16 in the label condition, and 10/16 in the no-label condition), and over 80% were reported to definitely understand at least seven of them (14/16 in the label condition, and 13/16 in the no-label condition). Finally, all children were reported to at least "maybe" understand all eight labels. Table 1 shows the percentage of children in each condition reported to "definitely" understand each label. This analysis confirms that the categories and labels used here were familiar.

Realistic, color drawings of a typical exemplar of each animal were obtained from commercially available picture books and were digitized for computer manipulation. These images will be referred to as the "demonstration animals." An additional typical exemplar of each category was created ("typical test animals"), primarily by manipulating the color of each standard. In addition, two hybrid animals were created for each set: The hybrids had features of both categories but were computer-generated to look more like a member of one category than the

TABLE 1
Study 1 Stimulus Sets

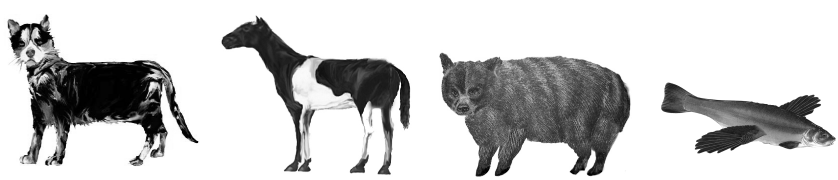
	Average Comprehension[a]		
Stimulus set	*Label*	*No-label*	*Activity (background photo in italics)*
Cat	94%	100%	Drinks *milk*
Dog	100%	100%	Eats *bones*
Horse	81%	81%	Sleeps in the *barn*
Cow	88%	94%	Sleeps outside on the *grass*
Bear	81%	88%	Lives in the *forest*
Pig	94%	81%	Lives in the *mud*
Fish	100%	94%	Lives in the *lake*
Bird	100%	100%	Lives in the *nest*

[a]Average percentage of participants in each condition reported by parents to "definitely" understand each animal label. Note that the percentages rise to 100% in each cell if the criterion is relaxed to "maybe" understands each label.

other. For example, in the cat-dog set, the hybrids had perceptual features of both cats and dogs, but one was designed to look more like a cat and the other was designed to look more like a dog. (Results from the no-label condition, described following, confirmed that children spontaneously treated these hybrids as members of the category they most resembled.) Children in the label condition heard these hybrids referred to with labels that did not match their appearances. Examples of the hybrid animals are shown in Figure 1. A set of warm-up stimuli was also used and consisted of pictures of typical dolls and shovels.

All images were sized to approximately 2 to 4 in. wide and 2 to 5 in. tall. Each image (and its left-right reverse) was printed in color, cut out, and mounted into a small stand so that it could stand on its own.

Each animal was associated with a particular activity, as shown in Table 1. Color photographs of a background or prop associated with each activity were ob-



Cat-like animal — Label: "Dog"
Horse-like animal — "Cow"
Bear-like animal — "Pig"
Fish-like animal — "Bird"

FIGURE 1 Examples of hybrid animals used in Study 1. Children saw the stimuli in color.

tained from a photo-object library (Hemera Technologies, Gatineau, Quebec, Canada). For example, the cat's activity was to drink milk, and the photo used to act out this activity was of a bowl of milk. These photos were mounted onto one of two 8.5 × 11-in. easels.

Procedure. Children were tested individually in the laboratory in a single session lasting about 15 min. They sat at a small table, with the researcher across from them. Each session began with a warm-up trial to introduce the task: On one easel, the experimenter displayed a photo of a bed and, using the doll stimulus, demonstrated and explained that a doll slept in the bed. This involved holding the doll up to the picture of a bed, rotating her as if she were lying down on it, and making snoring sounds. After this demonstration, the researcher handed the child the doll and encouraged him or her to "put the doll to sleep." Next, on the other easel, the researcher showed a photo of a bucket and, using the shovel stimulus, demonstrated and explained that a shovel could be used to scoop sand into or out of the bucket. This involved holding the shovel up to the picture of the bucket and making scooping motions, as if dropping sand into or scooping it out of the bucket. The researcher handed the child the shovel and encouraged him or her to use it to scoop sand himself or herself. Children were then shown additional doll and shovel exemplars in alternating order, and they were asked to show what each did, until they succeeded in placing a doll with the bed and a shovel with the bucket consecutively.

Test trials were similar to the warm-up trial. For example, the researcher displayed a photo of a bowl of milk on one easel and demonstrated and explained aloud that the demonstration cat drank milk. This involved holding the cat up to the picture of the bowl of milk and making sipping noises. The researcher then displayed a photo of dog bones on the other easel and demonstrated and explained aloud that the demonstration dog ate bones. This involved holding the dog up to the picture of the bones and making eating noises. As in the warm-up task, following each demonstration, children were handed the demonstration stimulus and were encouraged to act out the activity they had just seen demonstrated. To reduce memory load, the demonstration cat and dog were left standing in front of the photos of the bowl of milk and the bones, respectively. Children were then shown the three test exemplars (another typical cat, another typical dog, and either the cat-like hybrid or the dog-like hybrid), one at a time and in a pseudorandom order (described following). Their task was to act out, using the exemplar and the appropriate background photograph, whether each animal drank the milk or ate the bones. Half of the children in each condition saw the cat-like hybrid, and half saw the dog-like one.

One-half of the children participated in a label condition, in which they heard the experimenter use a category label when introducing each test animal (e.g., "Look at this dog! Can you show me what this dog does?"). The typical test animals were always called by their appropriate labels, and the hybrids were always called by labels that were the opposite of their appearance (e.g., the cat-like hybrid was referred to as "this dog;" see Figure 1). To establish baseline levels, the other half of the children participated in a no-label condition. They heard the experimenter use the phrase "this one" to introduce each test animal (e.g., "Look at this one! Can you show me what this one does?"). Regardless of children's selections, the researcher responded neutrally ("Thank you!"), and proceeded to the next test animal or the next set of animals.

The order in which the four animal sets were presented was counterbalanced according to a Latin Square design. The background photograph consistent with the activity associated with a hybrid's appearance appeared twice on the left and twice on the right for each child, and this was counterbalanced across children so that each photo appeared equally often on the left and right. Most children were presented with the four sets of animals shown in Table 1; however, due to fussiness, one child in the label condition did not complete the bird-fish set and one in the no-label condition did not complete the bear-pig set.

Coding of the children's responses was conducted off-line, via videotape, and involved noting which of the two background photos children indicated when presented with each test exemplar. Children often made the movements (and sounds if applicable) that had earlier been demonstrated, but pointing gestures were also coded as responses. Two coders, blind to condition, each coded one half of the sessions. To assess reliability, each coder also independently coded a random selection of one-fourth of the other coder's sessions; between-coder agreement was 99%.

Results and Discussion

Children's responses were scored as follows: A 1 was assigned if a child made a perceptually based inference by indicating the background photo associated with an animal's appearance (e.g., a cat or cat-like animal drank milk), and a 0 was assigned otherwise (e.g., a cat or cat-like animal ate bones). Preliminary analyses failed to reveal any effects or interactions involving sex; subsequent analyses collapsed across this factor.

Typical animals. To ensure that children understood the task, we first analyzed inferences children made about the typical animals. On seeing a typical cat, for example, children should infer that it would drink milk, rather than eat bones, regardless of whether they hear it referred to as "this cat" or "this one." Indeed, in the label and no-label conditions, children made perceptually based inferences about the typical items 79% (SD = 16%) and 70% (SD = 13%) of the time, respectively. These figures do not differ from each other, $t(30) = 1.62$, $p > .10$. Although children were not perfect in making the appropriate inferences about the typical items, they did so more frequently than expected by chance of 50%, *t*s > 6.3, *p*s < .001, suggesting that they understood the task.

Hybrid animals. The main question addressed by this study was whether toddlers would be willing to accept that an object that looked like a member of one category actually belonged to an entirely different category, just on the basis of hearing it referred to with the label of the nonobvious category. Table 2 shows the percentage of perceptually based inferences children made as a function of condition. The higher the number, the more likely children were to base their inferences on appearance.

As expected, children in the no-label condition, who heard the hybrids referred to neutrally, made perceptually based inferences 69% of the time, more frequently than expected by chance, $t(15) = 3.46$, $p < .01$. For example, on seeing the cat-like hybrid and hearing it referred to as "this one," children inferred that it drank milk rather than ate bones. In the absence of any label information, then, toddlers spontaneously considered the hybrids to be members of the category they most resembled. Indeed, they were as likely to consider the hybrids to be members of the category they most resembled as they were to consider the typical exemplars to be members of the category they most resembled.

Children in the label condition responded quite differently. When the hybrids were referred to with unexpected labels, children made perceptually based inferences only 37% of the time, marginally less frequently than expected by chance, $t(15) = 1.76$, $p < .10$. For example, when they heard the cat-like hybrid referred to as "this dog," children tended to infer that it ate bones rather than drank milk.

A 2 × 4 (Condition × Hybrid Animal) mixed Analysis of Variance showed that children in the no-label condition made significantly more perceptually based inferences than those in the label condition, $F(1, 28) = 15.40$, $p < .001$. There was no effect of hybrid, and there was no interaction, *F*s < 1, indicating that this pattern of results held across all four of the hybrid stimuli.

TABLE 2
Percentage of Perceptually Based Inferences About Hybrid Stimuli

	M	*SD*
Study 1 (Hybrid animals)		
No-label	69*	22
Label	37†	30
Study 2 (Hybrid artifacts)		
No-label	81*	22
Label	31*	22

*significantly different from chance of 50%, $p < .05$
†marginally below chance of 50%, $p < .10$

TABLE 3
Number of Children Making 0, 1, 2, 3, or 4 Perceptually Based Inferences About Hybrid Stimuli

	Number of Perceptually Based Inferences				
	0	*1*	*2*	*3*	*4*
Study 1: Hybrid animals					
No-label	0	1	3	8	3
Label	4	4	5	1	1
Study 2: Hybrid artifacts					
No-Label	0	0	3	3	6
Label	3	3	6	0	0

The difference between the no-label and label conditions was also evident at the individual level. The top portion of Table 3 shows the number of children who made 0, 1, 2, 3, or 4 perceptually based inferences about the hybrid animals. (The table and analysis do not include the child in the no-label condition or the one in the label condition who responded to only 3 of the 4 stimulus sets). Whereas 11 of the 15 (73%) children in the no-label condition made primarily perceptually based inferences (i.e., 3 or more), only 2 of the 15 (13%) in the label condition did so, distributions that are significantly different from each other, $\chi^2(1, N = 30) = 11.00$, $p < .001$.

Like the 2.5-year-olds in Gelman and Coley's (1990) study, even 24-month-olds recognized that the label applied to an object can be a better cue to category membership than appearance. They were willing to make an inference about an animal on the basis of the label they heard applied, even when that label conflicted with their own expectations about the category to which an animal belonged. In a second study, we sought to replicate these results using a different set of stimuli — namely, artifacts.

STUDY 2

Method

Participants. Twenty-four 24-month-olds (M = 24 months, 20 days; range = 23;1 to 26;3) participated. Twelve were girls, and 12 were boys. Seven additional toddlers were tested, but their data are not included due to extreme fussiness resulting in an inability to complete at least half of the session (3), responses solely to the left or right (2), parental over-involvement (1), or experimenter error (1). None of these children participated in Study 1.

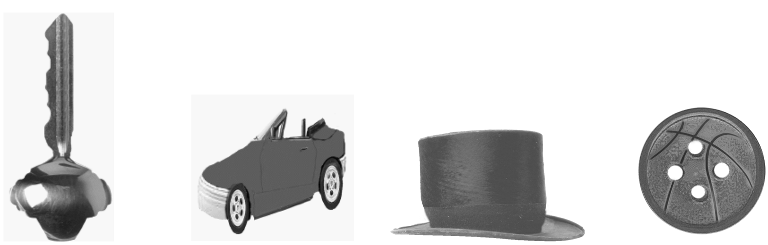
Design. Children were randomly assigned to a label or no-label condition, resulting in 12 participants per condition, balanced for gender. The average age in the label condition was 24;19 and in the no-label condition 24;20.

Materials. Eight artifacts from familiar categories were grouped into four pairs based roughly on shape similarity: key-spoon, shoe-car, cup-hat, and button-ball. As in Study 1, familiarity was confirmed by a brief vocabulary checklist that asked parents to indicate if their children definitely understood, maybe understood, or did not understand the eight category labels. At least 58% of the children in each condition definitely understood all eight labels (10/12 in the label condition and 7/12 in the no-label condition), and all children in both conditions definitely understood at least seven of them. Finally, all children were reported to at least "maybe" understand all eight labels. Table 4 shows the percentage of children in each condition reported to "definitely" understand each label. This analysis confirms that the categories and labels used here were familiar.

Color photographs of a typical exemplar of each artifact were obtained from a digital library of photo-objects (Hemera Technologies, Gatineau, Quebec, Canada). Typical and hybrid test artifacts were created in the same manner as described in Study 1. Examples of the artifact hybrids are shown in Figure 2.

Each artifact was associated with a typical function, as shown in Table 4. As in Study 1, color photographs mounted onto easels were used to demonstrate those functions (see Procedure).

Procedure. The procedure was the same as that used in Study 1. For example, the researcher used photographs of a cereal bowl and a car to demonstrate that the spoon could be used to eat cereal and the key could be used to start the car. As in Study 1, to eliminate any memory requirement for the associations during the test trials, the standard spoon remained in front of the bowl and the standard key re-



Key-like artifact	Car-like artifact	Hat-like artifact	Button-like artifact
Label: "Spoon"	"Shoe"	"Cup"	"Ball"

FIGURE 2 Examples of hybrid artifacts used in Study 2. Children saw the stimuli in color.

TABLE 4
Study 2 Stimulus Sets

	Average Comprehension[a]		
Stimulus set	*Label*	*No-label*	*Function (background photo in italics)*
Key	100%	92%	Starts the *car*
Spoon	92%	100%	Used to eat cereal from the *bowl*
Shoe	100%	100%	Goes on the *baby's* foot
Car	100%	100%	Drives on the *road*
Cup	100%	100%	Sits on the *table*
Hat	100%	100%	The *man* wears it
Button	92%	67%	Goes on the *coat*
Ball	100%	100%	Drops through the *hoop*

[a]Average percentage of participants in each condition reported by parents to "definitely" understand each artifact label. Note that the percentages rise to 100% in each cell if the criterion is relaxed to "maybe" understands each label.

mained in front of the car. Children were asked to demonstrate the function associated with the typical and hybrid test items. As before, test items were either introduced with labels or not: When labeled, typical artifacts were always called by their appropriate names, and hybrids were always referred to with labels opposite to their appearance (e.g., the key-like object was referred to as "this spoon").

Coding of responses was conducted off-line, via videotape, and involved noting which of the two photographs on each trial children used or indicated when demonstrating an artifact's function. As in Study 1, two coders, blind to condition, each coded one half of the sessions, and each also independently coded a random one-fourth of the other's. Agreement was 97%.

Results and Discussion

Children's responses were scored as follows: A 1 was assigned if a child made a perceptually based inference about an artifact's function (e.g., used a key or key-like object to start the car), and a 0 was assigned otherwise (e.g., used a key or key-like object to eat cereal from a bowl). Preliminary analyses failed to reveal any effects or interactions with sex; subsequent analyses collapsed across this factor.

Typical artifacts. In both conditions, children should have made primarily perceptually based inferences about the typical artifacts: In the label condition, they heard them referred to with labels that match their appearance, and in the no-label condition, they heard them referred to neutrally. Indeed, children in the label and no-label conditions made perceptually based inferences 83% (SD = 25%)

and 93% (SD = 15%) of the time, respectively. The two conditions did not differ from each other, $t(22) = 1.26$, $p > .10$, and both represent more perceptually based inferences than would be expected by chance, ts > 4.69, ps < .01. As in Study 1, then, children in both conditions clearly understood the task.

Hybrid artifacts. The percentage of perceptually based inferences made about the hybrid artifacts is shown in Table 2. Consistent with Study 1, children in the no-label condition treated the hybrid items like members of the categories they most resembled. They made perceptually based inferences 81% of the time, more frequently than chance, $t(11) = 5.00$, $p < .01$. For example, when they heard the key-like hybrid referred to as "this one," they tended to use it to start the car (rather than to eat cereal from a bowl). When children heard the hybrids referred to with labels that did not match their appearances, however, they made perceptually based inferences just 31% of the time, significantly less frequently than expected by chance, $t(11) = 3.00$, $p < .05$. When they heard the key-like object referred to as "this spoon," they tended to use it to eat cereal (rather than start the car).

As in Study 1, a 2 × 4 (Condition × Hybrid Artifact) mixed Analysis of Variance showed that children in the label condition made significantly fewer perceptually based inferences than those in the no-label condition, $F(1, 22) = 32.00$, $p < .001$. There was no effect of hybrid, and there was no interaction, Fs < 1, indicating that this pattern of results held across all four of the hybrid stimuli.

The lower half of Table 3 shows the response patterns by individual. As the table shows, 9 of the 12 (75%) children in the no-label condition made predominantly perceptually based inferences (i.e., 3 or more), whereas none of the 12 children in the label condition did so, $\chi^2(1, N = 24) = 14.4$, $p < .001$.

GENERAL DISCUSSION

These two studies show that 24-month-olds are willing to give up a compelling, perceptually based classification in favor of an unexplained, and potentially baffling, label-based one. In both the animal and artifact domains, labeling profoundly affected the inferences children made. For example, in Study 2, children who heard a key-like object referred to neutrally inferred that it could be used to start a car; those who heard the same object referred to as a "spoon" instead inferred that it was used to eat cereal from a bowl.

This work can be viewed as part of the tradition in developmental research emphasizing the social aspects of knowledge acquisition (e.g., Bruner & Haste, 1987; Rogoff, 1990; Tomasello, 1999; Vygotsky, 1978). As a number of researchers have pointed out, the ways in which caregivers talk about the world can contribute to the theories children form. For example, Jipson and Callanan (2003) found that when looking through a picture book depicting various objects increasing in size, mothers of 3- and 4-year-olds were much more likely to use the word grow when commenting on biological kinds (e.g., a mushroom getting larger) than artifact kinds (e.g., a beach ball expanding). Furthermore, they were also more likely to explain changes in the size of biological kinds by referring to an organic cause and changes in the size of artifact kinds by referring to a human cause. As Jipson and Callanan argue, this kind of discourse may help children to divide the world into distinct domains and to reason about those domains appropriately.

The explanations and elaborations parents use when introducing labels can also affect the categories children form. For example, on the basis of simple ostension (e.g., "This is an *animal*"), children sometimes fail to learn a second term for an object they already have a label for (e.g., Jaswal & Hansen, 2006; Liittschwager & Markman, 1994; Macnamara, 1982; Merriman, 1986). However, when a speaker provides testimony describing the relation between the two terms (e.g., "A dog is a kind of *animal*"), children more readily accept, learn, and extend that category term (e.g., Callanan, 1989, 1991; Clark, 1997; Clark & Grossman, 1998). Similarly, Jaswal (2004) found that 4-year-olds who reacted skeptically to hearing a key-like object called a spoon, for example, were much more likely to accept that it was, in fact, a spoon if the speaker testified to the fact that the label was surprising (e.g., "You're not going to believe this, but this is actually a spoon"). In short, children can learn the conventional ways to reason about objects, events, and people by paying attention to how others talk about them and by being receptive to that testimony (see also Astuti, Solomon, & Carey, 2004; Callanan & Oakes, 1992; Gelman, Taylor, & Nguyen, 2004).

The two studies reported here have shown that labeling itself, without any additional explanation or elaboration, can serve as an important form of testimony. These results are important for several reasons. First, although Graham et al. (2004) and Waxman and Markow (1995) have shown that labeling can influence the categories that even 12- and 13-month-olds form, their studies focused on how a novel label can lead children to form a new category. Our studies, in contrast, asked if a familiar label could lead children to recategorize an object from one known category to another known category. It is important to emphasize that the hybrids in our studies were not ambiguous in appearance; each one was designed to look more like one familiar category than the other, and results from the no-label conditions confirmed that children spontaneously considered them to be members of the familiar category they most resembled (see also Jaswal, 2004). This represents an especially strong test of the power of a label to influence categorization because making use of the object's label when drawing an inference required children to "disbelieve their eyes."

Gelman and Coley (1990) showed that 2.5-year-olds drew inferences about an object based on a familiar category label, even when it conflicted with the object's appearance, but this question had not yet been addressed with younger children. Indeed, there was some reason to suspect that younger children would behave differently. For example, Banigan and Mervis (1988) found that 24-month-olds rarely reclassified a miscategorized object on the basis of merely hearing it referred to with an unexpected label. However, their measure was children's later comprehension and production, or both, of the label. Poor performance could have been the result either of a resistance to accepting the unexpected classification in the first place, or of a difficulty in remembering how it applied later. Our studies used an inference task that did not require verbal responses and that minimized memory demands. Results showed clearly that 24-month-olds are, in fact, receptive to accepting a label-based classification, even if it conflicts with a perceptually based one.

One might wonder whether children are always receptive to anomalous category information. Using an inference procedure similar to the one used here, Jaswal (2004) recently showed that 4-year-olds were often quite skeptical about unexpected labels they heard another person use. They often interpreted the labels as errors on the speaker's part — the result of ignorance or incomplete perceptual access, for example. Younger children may be more receptive to anomalous category information because they have less experience with even the familiar categories used here (see also Naigles, Gleitman, & Gleitman, 1992, for an analogous argument in verb subcategorization).

Another reason these results are important is that they add to a growing body of work suggesting that even very young children's categories are not perceptually based at their core (for reviews, see Bloom, 2000; Gelman, 2003). Although the vast majority of category labels in a typical 2-year-old's vocabulary do represent categories whose members cohere perceptually (Samuelson & Smith, 1999), children this age nonetheless seem to expect that something deeper than surface similarity is responsible for category membership. In fact, the willingness of 24-month-olds to make an inference on the basis of an object's category label, even when that label conflicts with its appearance, suggests that they recognize that the appearance of an object can actually be misleading. An explicit understanding of the appearance-reality distinction may not develop until about 4.5 years of age (e.g., Flavell, Flavell, & Green, 1983; Flavell, Green, & Flavell, 1986; but see Hansen & Markman, 2005), but an implicit understanding that things are not always what they seem is available considerably earlier.

Finally, as noted earlier, these results are important also because they emphasize the influence of testimony on children's early categories. Consider how we know whether whales are mammals or fish. On the surface, whales seem to share more in common with fish: They live in water, they have fins, they swim, and so on. Despite these superficial similarities with fish, however, taxonomists classify whales as mammals. This classification is hardly controversial, even though it requires us to "disbelieve our eyes," and even though most of us would never have the time or specialized knowledge to verify it ourselves.

In everyday life, it may not matter whether a whale is mammal or a fish. However, when we want to make predictions about something or explain its behavior, knowing the category in which it belongs is very important. A willingness to accept the classification an expert provides enables a learner to take advantage of that person's knowledge. Having a reliable source explain that whales are mammals allows us to infer that whales bear live young, breathe air, have warm blood, and so on — inferences that are quite different from those we would make if we classified whales as fish.

A willingness to accept testimony on matters of categorization seems related to an assumption adults make when communicating, which Putnam (1973) called the "division of linguistic labor." According to Putnam, adults routinely use terms without knowing the criteria for their use. For example, we readily use the word "gold" to refer to the precious metal, even though very few of us actually know how to distinguish real gold from fool's gold. Nonetheless, Putnam argued, we assume that there are criteria and that knowledge of the criteria for category membership is possessed by at least some members of our community.

Similarly, in the studies reported here, 24-month-olds were willing to set aside a spontaneously generated classification in favor of a classification they could not have immediately understood — apparently accepting "on faith" that there was a reason why something that looked very much like a key, for example, was being called a "spoon." Furthermore, toddlers were willing to make this leap of faith, even though the source of the labels was an unfamiliar adult speaker whom they had just met. This suggests that, by default, they may be willing to accept testimony, even when that testimony is unexpected. Recent work by Koenig, Clement, and Harris (2004) has shown that older children will discount information from an adult who has been unreliable in the past. It remains to be seen whether toddlers would also react more skeptically to unexpected information that came from a previously unreliable source.

The studies presented here were not designed to address the limits of children's receptiveness to unexpected labels — for example, whether they would be willing to accept that a cat-like animal was a "fish," or that a key-like object was a "basketball." When these kinds of obvious misnaming events do occur, previous research has suggested that even very young children react skeptically. For example, Koenig and Echols (2003) reported that most 16-month-olds spontaneously attempted to correct a speaker who, for example, referred to a shoe as a "ball" (see also Pea, 1982). Similarly, Graham et al. (2004, Experiment 3) found that when 13-month-olds heard two similar-looking novel objects referred to with different labels, they nonetheless treated the two objects as members of the same category — essentially ignoring the fact that they had been called by different names. In this case, the perceptual similarity of the two objects appears to have been too salient for the different labels to have an effect.

The hybrids used in our studies were designed to be unlikely exemplars of the named category (although they did share its overall shape), were from the same ontological category, and even had a few of its characteristics (e.g., the button-like object, which was referred to as a "ball," had the coloring of a basketball). Nevertheless, children spontaneously considered each one to be a member of the familiar category it most resembled, and hearing an alternative familiar label led them to treat it like a member of an entirely different familiar category.

The power of a simple label to convey unexpected category information may well develop with age and experience with language (e.g., Gentner & Rattermann, 1991; Mareschal, 2003; Sloutsky, Lo, & Fisher, 2001). Certainly by 24 months, however, this process is in place, and it can enable children to take advantage of the rich and conceptual frameworks that their cultures have evolved.

ACKNOWLEDGMENTS

Generous support for this research was provided by the University of Virginia.

Portions of this work were presented at the International Conference on Infant Studies, April, 2002, in Toronto, and the Cognitive Science Society Meeting, August, 2002, in Fairfax, VA.

We would like to thank the children and parents who participated. We also thank Keelah Andrews, Olivia Lima, Lina Liu, Lauren Malone, Brittany McLaughlin, Yujin Moon, and Leslie Neely for assistance in data collection and coding.

REFERENCES

Adams, A. K., & Bullock, D. (1986). Apprenticeship in word use: Social convergence processes in learning categorically related nouns. In S. A. Kuczaj & M. D. Barrett (Eds.), *The development of meaning* (pp. 155–197). New York: Springer-Verlag.

Astuti, R., Solomon, G. E. A., & Carey, S. (2004). Constraints on conceptual development. *Monographs of the Society for Research in Child Development, 69 (3, Serial No. 277).*

Baldwin, D. A., Markman, E. M., & Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development, 64*, 711–728.

Banigan, R. L., & Mervis, C. B. (1988). Role of adult input in young children's category evolution: II. An experimental study. *Journal of Child Language, 15*, 493–504.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Bruner, J., & Haste, H. (1987). *Making sense: The child's construction of the world*. London: Methuen.

Callanan, M. A. (1989). Development of object categories and inclusion relations: Preschoolers' hypotheses about word meanings. *Developmental Psychology, 25*, 207–216.

Callanan, M. A. (1991). Parent-child collaboration in young children's understanding of category hierarchies. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development* (pp. 440–484). Cambridge: Cambridge University Press.

Callanan, M. A., & Oakes, L. A. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development, 7*, 213–233.

Clark, E. V. (1997). Conceptual perspective and lexical choice in acquisition. *Cognition, 64*, 1–37.

Clark, E. V., & Grossman, J. B. (1998). Pragmatic directions and children's word learning. *Journal of Child Language, 25*, 1–18.

Coady, C. A. J. (1992). *Testimony: A philosophical study*. Oxford, England: Oxford University Press.

Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Development of the appearance-reality distinction. *Cognitive Psychology, 15*, 95–120.

Flavell, J. H., Green, F. L., & Flavell, E. R. (1986). Development of knowledge about the appearance-reality distinction. *Monographs of the Society for Research in Child Development, 51* (Serial No. 212), 1–68.

Gelman, S. A. (2003). *The essential child*. Oxford, England: Oxford University Press.

Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology, 26*, 796–804.

Gelman, S. A., Collman, P., & Maccoby, E. E. (1986). Inferring properties from categories versus inferring categories from properties: The case of gender. *Child Development, 57*, 396–404.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*, 183–208.

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development, 58*, 1532–1541.

Gelman, S. A., Taylor, M. G., & Nguyen, S. P. (2004). Mother-child conversations about gender. *Monographs of the Society for Research in Child Development, 69 (1, Serial No. 275).*

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development* (pp. 225–277). Cambridge: Cambridge University Press.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71*, 1205–1222.

Graham, S. A., Kilbreath, C. S., & Welder, A. N. (2004). Thirteen-month-olds rely on shared labels and shape similarity for inductive inferences. *Child Development, 75*, 409–427.

Hansen, M. B., & Markman, E. M. (2005). Appearance questions can be misleading: A discourse-based account of the appearance-reality problem. *Cognitive Psychology, 50*, 233–263.

Harris, P. L. (2002). What do children learn from testimony? In P. Carruthers, S. Stitch, & M. Siegal (Eds.), *The cognitive basis of science*. Cambridge, UK: Cambridge University Press.

Jaswal, V. K. (2004). Don't believe everything you hear: Preschoolers' sensitivity to speaker intent in category induction. *Child Development, 75*, 1871–1885.

Jaswal, V. K., & Hansen, M. B. (2006). Learning words: Children disregard some pragmatic information that conflicts with mutual exclusivity. *Developmental Science, 9,* 158–165.

Jipson, J. L., & Callanan, M. A. (2003). Mother-child conversation and children's understanding of biological and nonbiological changes in size. *Child Development, 74*, 629–644.

Jones, S. S., & Smith, L. B. (1993). The place of perception in children's concepts. *Cognitive Development, 8*, 113–139.

Kemler Nelson, D. G., Russel, R., Duke, N., & Jones, K. (2000). Two-year-olds will name artifacts by their functions. *Child Development, 71*, 1271–1288.

Koenig, M. A., Clement, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science, 15*, 694–698.

Koenig, M. A., & Echols, C. H. (2003). Infants' understanding of false labeling events: The referential roles of words and the speakers who use them. *Cognition, 87*, 179–208.

Liittschwager, J. C., & Markman, E. M. (1994). Sixteen- and 24-month-olds' use of mutual exclusivity as a default assumption in second-label learning. *Developmental Psychology, 30*, 955–968.

Macnamara, J. (1982). *Names for things*. Cambridge, MA: MIT Press.

Mareschal, D. (2003). The acquisition and use of implicit categories in early development. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming, buzzing confusion* (pp. 360–383). Oxford, England: Oxford University Press.

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

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology, 16*, 1–27.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge: Cambridge University Press.
Meints, K., Plunkett, K., & Harris, P. L. (1999). When does an ostrich become a bird? The role of typicality in early word comprehension. *Developmental Psychology, 35*, 1072–1078.
Merriman, W. E. (1986). Some reasons of the occurrence and eventual correction of children's naming errors. *Child Development, 57*, 942–952.
Mervis, C. B. (1984). Early lexical development: The contributions of mother and child. In C. Sophian (Ed.), *Origins of cognitive skills* (pp. 339–370). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Mervis, C. B. (1987). Child-basic object categories and early lexical development. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 201–233). Cambridge: Cambridge University Press.
Mervis, C. B., Pani, J. R., & Pani, J. R. (2003). Transaction of child cognitive-linguistic abilities and adult input in the acquisition of lexical categories at the basic and subordinate levels. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming, buzzing confusion* (pp. 242–274). Oxford, England: Oxford University Press.
Naigles, L. G., Gleitman, H., & Gleitman, L. R. (1992). Children acquire word meaning components from syntactic evidence. In E. Dromi (Ed.), *Language and cognition: A developmental perspective* (pp. 104–140). Norwood, NJ: Ablex.
Nazzi, T., & Gopnik, A. (2001). Linguistic and cognitive abilities in infancy: When does language become a tool for categorization? *Cognition, 80*, B11–B20.
Pea, R. D. (1982). Origins of verbal logic: Spontaneous denials by two- and three-year-olds. *Journal of Child Language, 9*, 597–626.
Putnam, H. (1973). Meaning and reference. *The Journal of Philosophy, 70*, 699–711.
Quinn, P. C., Eimas, P. D., Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception, 22*, 463–475.
Rakison, D. H., & Oakes, L. M. (2003). *Early category and concept development: Making sense of the blooming, buzzing confusion*. Oxford, England: Oxford University Press.
Reid, T. (1764/1997). *An inquiry into the human mind on the principles of common sense*. University Park: Pennsylvania State University.
Rogoff, B. (1990). *Apprenticeship in thinking*. New York: Oxford University Press.
Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382–439.
Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition, 73*, 1–33.
Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in children: A similarity-based model. *Journal of Experimental Psychology: General, 133*, 166–188.
Sloutsky, V. M., Lo, Y., & Fisher, A. V. (2001). How much does a shared name make things similar: Linguistic labels, similarity, and the development of inductive inference. *Child Development, 72*, 1695–1709.
Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.
Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*, 257–302.
Welder, A. N., & Graham, S. A. (2001). The influence of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development, 72*, 1653–1673.