Journal of Experimental Child Psychology 114 (2013) 432–455

Contents lists available at SciVerse ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

The role of linguistic labels in inductive generalization

W. Deng, Vladimir M. Sloutsky *

Department of Psychology and Center for Cognitive Science, The Ohio State University, Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Received 20 June 2012
Revised 31 August 2012
Available online 25 December 2012

Keywords:
Categorization
Category learning
Inductive inference
Conceptual development
Attention
Cognitive development

ABSTRACT

What is the role of linguistic labels in inductive generalization? According to one approach labels denote categories and differ from object features, whereas according to another approach labels start out as features and may become category markers in the course of development. This issue was addressed in four experiments with 4- and 5-year-olds and adults. In Experiments 1 to 3, we replicated Yamauchi and Markman's findings with adults (*Journal of Memory and Language,* 1998, Vol. 39, pp. 124–148, and *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 2000, Vol. 26, pp. 776–795) and extended the paradigm to young children. In Experiment 4, we compared effects of labels with those of highly salient visual features. Overall, results of these experiments provide strong support for the idea that early in development labels function the same way as other features, but they may become category markers in the course of development. A related finding is that whereas categorization and induction may be different processes in adults, they seem to be equivalent in young children. These results are discussed with respect to theories of development of inductive generalization.

Introduction

Induction, or generalizing knowledge from known to novel, is a critical component of learning and cognition; induction enables us to apply learned knowledge to new situations. Some examples of inductive generalization include (a) inferring a property of a novel item given that a known item has this property and (b) inferring a category of a novel item given category membership of a known item. The former is referred to as *projective induction*, and the latter is referred to as *categorization.* The term *induction* is often used to refer to both projective induction and categorization (Sloutsky & Fisher, 2004a).

* Corresponding author.
E-mail address: sloutsky.1@osu.edu (V.M. Sloutsky).

0022-0965/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jecp.2012.10.011

Induction may have humble beginnings; it has been well established that induction appears early in development (Gelman & Markman, 1986; Mandler & McDonough, 1996; Sloutsky & Fisher, 2004a; Sloutsky & Fisher, 2008). There is also much evidence demonstrating that even early in development linguistic labels may affect inductive generalization (Gelman & Markman, 1986; Sloutsky & Fisher, 2004a; Sloutsky, Lo, & Fisher, 2001; Welder & Graham, 2001). However, the mechanism underlying the role of labels in early induction is hotly debated. Do labels start out as category markers (i.e., symbols denoting the category), or do they start out as features and potentially become category markers in the course of development. In what follows, we consider both possibilities in greater detail.

Putative mechanisms underlying effects of labels on generalization

Some researchers have argued that from early in development, children expect linguistic labels (primarily in the form of count nouns) to mark categories (Waxman & Markow, 1995) and facilitate inductive generalization (e.g., Gelman, 2003; Welder & Graham, 2001). According to this view, a common label suggests a common category (e.g., if two items are called "dog," then they are likely to belong to the same kind), whereas a common category suggests that the items may share multiple properties. Therefore, when performing induction, people may first use a category label to identify the category to which the entity belongs and then generalize properties of that entity to other members of the target category. For example, in a series of experiments, Gelman and Markman (1986) presented young children with triads consisting of a target and two test items. One test item shared the label with the target but looked dissimilar from it, whereas the other test item looked similar to the target but had a different label. Children were informed that one test item had a particular hidden property (e.g., "hollow bones") and the other test item had a different hidden property (e.g., "solid bones"), and they were asked to decide which hidden property the target had. The results indicated that children were more likely to base their inference on the common label than on perceptual similarity (but see Sloutsky & Fisher, 2004a, Experiment 4, for diverging evidence and counterarguments). This and similar findings have been interpreted as evidence that children's induction is based on category membership, which is denoted by a particular label.

There is also evidence that count nouns are more likely to guide induction than other word forms. For example, Gelman and Heyman (1999) reported that young children were more willing to generalize properties of a person from one context to another when the person was referred to by a count noun (e.g., "carrot-eater") than when the person was referred to by a descriptive sentence (e.g., "likes to eat carrots").

These findings, however, do not lend unequivocal support to the idea that words are category markers. For example, some researchers have suggested that the contribution of linguistic labels is driven by attentional rather than conceptual factors (Napolitano & Sloutsky, 2004; Sloutsky & Napolitano, 2003). There is also evidence that labels contribute to the overall similarity of compared entities (Sloutsky & Fisher, 2004a; Sloutsky & Lo, 1999) and, thus, to both categorization and induction. In one experiment using items that had been previously used by Gelman and Markman (1986), Sloutsky and Fisher (2004a) demonstrated that similarity computed over labels and appearances can accurately predict young children's responses, whereas a model that assumes reliance only on labels fails to predict children's performance. Proponents of this view have also argued that early in development labels may function like other features (e.g., shape, color, size), although they may become category markers as a result of development (Deng & Sloutsky, 2012; Sloutsky, 2010; Sloutsky & Fisher, 2004a; Sloutsky & Lo, 1999; Sloutsky et al., 2001).

In short, according to one approach, labels start out as category markers; even early in development they denote categories and, as such, differ from other features. In contrast, labels may become category markers as a result of development, whereas early in development labels do not qualitatively differ from other features.

Experimental distinction between labels-as-features and labels-as-category-markers

In an attempt to distinguish between labels being features and category markers, Yamauchi and Markman (1998, 2000) developed an innovative paradigm potentially capable of settling the issue.

The paradigm is based on the following idea. Imagine two categories, labeled "A" and "B", each having four binary dimensions (e.g., size: large vs. small; color: black vs. white; shape: square vs. circle; texture: smooth vs. rough). The prototype of Category A has all values denoted by "1" (i.e., A, 1, 1, 1, 1), and the prototype of Category B has all values denoted by "0" (i.e., B, 0, 0, 0, 0). There are two interrelated generalization tasks: categorization (referred to as "classification" by the authors) and projective induction (referred to as "inference"). The goal of classification is to predict category membership (and hence the label) on the basis of presented features. For example, participants are presented with all of the values for an item (e.g., ?, 0, 1, 1, 1) and need to predict category label A or B. In contrast, the goal of inference is to predict a feature on the basis of category label and other presented features. For example, given an item (e.g., A, 1, ?, 1, 0), participants need to predict the value of the missing feature. A critical manipulation that could illuminate the role of labels is the "low-match" condition. For low-match inference, participants were presented with an item "A" (?, 0, 1, 0, 0, which had more features in common with the prototype of Category B but had label A) and asked to predict the missing feature. For low-match classification, participants were presented with an item "?" (1, 0, 1, 0, 0, which again had more features in common with the prototype of Category B) and asked to predict the missing label.

Yamauchi and Markman (1998, 2000) reasoned that if the label is just a feature, then performance on the low-match classification and inference tasks should be symmetrical. However, if labels are more than features and are treated as category markers, then predicting a label when features are provided (i.e., a classification task) should elicit different performance from a task of predicting a feature when the label is provided (i.e., an inference task). Specifically, category-consistent responding should be more likely in low-match inference tasks (where participants can rely on the category label) than in low-match classification tasks (where participants need to infer the category label).

On finding predicted asymmetries between the two conditions, Yamauchi and Markman (1998, 2000) concluded that category labels differed from other features (see also Rehder, Colner, & Hoffman, 2009, for supporting eye-tracking evidence). These findings have been replicated in a series of follow-up studies (see Markman & Ross, 2003, for a review) and have been successfully modeled (see Love, Medin, & Gureckis, 2004).

What is at stake: why is the difference between labels-as-features and labels-as-category-markers important?

Why is understanding the role of label early in development important? We believe that there are at least two reasons. First, this understanding is necessary for identifying the mechanism of early generalization and its change in the course of development, and this knowledge in turn may elucidate more general principles of cognitive development. In particular, if labels function as features, they contribute to generalization in a bottom-up manner (by contributing to the featural overlap among the compared items), whereas if they are category markers, they may guide the process in a top-down manner (by triggering a search for overlapping features). Each of these possibilities has long-ranging consequences for our understanding of cognitive development. If from early in development language exerts top-down influences on category learning, then even early in development the lower level processes (e.g., discrimination, generalization) are subject to top-down control. Therefore, the ability to exert top-down control, as well as cognitive and neural mechanisms that subserve this ability, need to exhibit early onset. Alternatively, if words acquire the ability to guide cognition in the course of development, then top-down control does not need to exhibit early onset and could itself be a product of development.

Second, the role that labels play in generalization may elucidate relationships between categorization and induction. Note that some researchers have argued that that the two tasks are functionally equivalent for adults (e.g., Anderson, 1991) and children (e.g., Sloutsky & Fisher, 2004a), whereas others have argued that the tasks are functionally different (see Markman & Ross, 2003, for a review). If the tasks are equivalent, then the representation formed in the course of classification and inference training should be equivalent as well. In contrast, if the tasks are functionally different, then classification and inference training should result in different representations. For example, Markman and Ross (2003) presented an extensive argument regarding potential differences in representations

between classification and inference training and presented evidence supporting this distinction in adults (see also Hoffman & Rehder, 2010, for eye-tracking evidence, and Love et al., 2004, for a computational model). However, if early in development labels function as features, then classification and inference tasks should be equivalent, which in turn suggests that extensive differences observed between classification and inference in adults are a product of development. We return to this issue in the General Discussion section.

The current research

Yamauchi and Markman's (1998, 2000) paradigm has been successfully applied for examining the role of labels in adults' generalization and could be applied for examining possible developmental changes in the effect of labels on generalization. Does the asymmetry between low-match classification and low-match inference characterizing adults' performance also characterize children's performance? Finding such an asymmetry would suggest that labels play a similar role across development, indicating that even for young children labels are more than features. However, as argued above, it is possible that labels function differently across development; whereas labels may denote function as category markers in adults, they may function as perceptual features in young children. If this is the case, then unlike adults, children may exhibit symmetrical performance in low-match classification and low-match inference.

The reported experiments were designed to address these issues. In Experiments 1 to 3, we replicated Yamauchi and Markman's (2000) findings with adults and extended the paradigm to young children. In Experiment 4, we compared effects of labels with those of highly salient visual features.

Experiment 1

The goal of Experiment 1 was to (a) replicate Yamauchi and Markman's (2000) paradigm with adults and (b) examine the role of labels in early generalization by extending the paradigm to young children. Similar to Yamauchi and Markman, participants learned two categories of creatures and then were given classification and inference trials, half of which were high-match and half of which were low-match. There were small procedural differences between the current procedure and the one used by Yamauchi and Markman. Most important, in contrast to Yamauchi and Markman, where labels were presented as a single written word, labels in Experiment 1 were presented auditorily in a carrier phrase (e.g., "This is a Flurp").

Based on Yamauchi and Markman's (2000) results, we expected that adult participants would make category-consistent responses in low-match inference but not in low-match classification. This finding would be consistent with the idea that adults treat labels as category markers. Finding such an asymmetry in young children would support the idea that even for children labels are more than features.

Experiment 1A

Method

Participants

Participants were 12 adults (3 women and 9 men) and 12 preschool children (mean age = 56.0 - months, range = 51.9–59.4, 7 girls and 5 boys). In this and all other experiments reported here, children were recruited from child care centers located in middle-class suburbs of Columbus, Ohio, in the U.S. Midwest, and tested in a quiet room in their preschool by a female experimenter. All adults were undergraduate students from The Ohio State University participating for course credit.

Materials

In all reported experiments, materials were colorful drawings of artificial creatures measuring 17.0 by 23.5 cm (see Fig. 1). The items had five features varying in color and shape and formed two

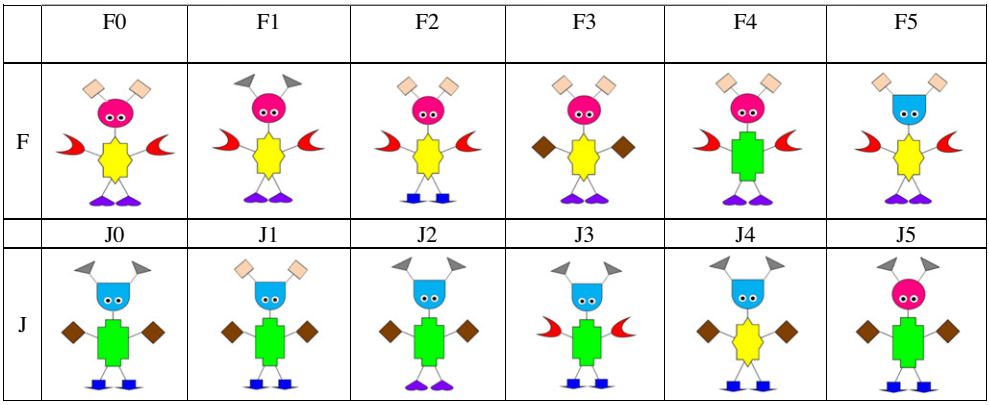

Fig. 1. Stimuli examples from two categories used in Experiments 1 to 3. F, Flurp; J, Jalet. F0 and J0 are prototypes of each category, and F1/J1–F5/J5 are individual exemplars.

categories determined by feature values. Artificial labels ("Flurp" or "Jalet" printed above each creature) were used to refer to the categories.

As shown in Tables 1, 2A and 2B), the two categories have a family resemblance structure that is derived from two prototypes (F0 and J0) by modifying the values of one of five features (see Fig. 1).

Table 1
Category structure used in learning in Experiments 1 to 4.

Flurp							Jalet						
Stimuli	Head	Body	Hands	Feet	Antenna	Label	Stimuli	Head	Body	Hands	Feet	Antenna	Label
F1	1	1	1	1	0	1	J1	0	0	0	0	1	0
F2	1	1	1	0	1	1	J2	0	0	0	1	0	0
F3	1	1	0	1	1	1	J3	0	0	1	0	0	0
F4	1	0	1	1	1	1	J4	0	1	0	0	0	0
F5	0	1	1	1	1	1	J5	1	0	0	0	0	0
F0	1	1	1	1	1	1	J0	0	0	0	0	0	0

Note. The value 1 = any of five dimensions identical to "Flurp" (see Fig. 1). The value 0 = any of five dimensions identical to "Jalet" (see Fig. 1). F, Flurp; J, Jalet. F0 and J0 are prototypes of each category.

Table 2A
Structure of testing stimuli in classification used in Experiments 1 to 3.

Flurp							Match	Jalet						
Stimuli	Head	Body	Hand	Feet	Antenna	Target label		Stimuli	Head	Body	Hand	Feet	Antenna	Target label
F11	1	1	1	1	0	?	High	J11	0	0	0	0	1	?
F12	1	1	1	0	1	?		J12	0	0	0	1	0	?
F13	1	1	0	1	1	?		J13	0	0	1	0	0	?
F14	1	0	1	1	1	?		J14	0	1	0	0	0	?
F15	0	1	1	1	1	?		J15	1	0	0	0	0	?
F21	1	0	1	0	0	?	Low	J21	0	1	0	1	1	?
F22	0	1	0	1	0	?		J22	1	0	1	0	1	?
F23	0	0	1	0	1	?		J23	1	1	0	1	0	?
F24	1	0	0	1	0	?		J24	0	1	1	0	1	?
F25	0	1	0	0	1	?		J25	1	0	1	1	0	?

Note. High and low are two levels of feature match. F, Flurp; J, Jalet. Category-consistent responses were the ones consistent with the values indicated in the target features and target labels.

Table 2B
Structure of testing stimuli in induction in Experiments 1 to 3.

Flurp							Match	Jalet						
Stimuli	Head	Body	Hand	Feet	Antenna	Target label		Stimuli	Head	Body	Hand	Feet	Antenna	Target label
F11	1	?	1	1	0	1	High	J11	0	?	0	0	1	0
F12	1	1	?	0	1	1		J12	0	0	?	1	0	0
F13	?	1	0	1	1	1		J13	?	0	1	0	0	0
F14	1	0	1	1	?	1		J14	0	1	0	0	?	0
F15	0	1	1	?	1	1		J15	1	0	0	?	0	0
F21	?	0	1	0	0	1	Low	J21	?	1	0	1	1	0
F22	0	?	0	1	0	1		J22	1	?	1	0	1	0
F23	0	0	?	0	1	1		J23	1	1	?	1	0	0
F24	1	0	0	?	0	1		J24	0	1	1	?	1	0
F25	0	1	0	0	?	1		J25	1	0	1	1	?	0

Note. High and low are two levels of feature match. F, Flurp; J, Jalet. Category-consistent responses were the ones consistent with the values indicated in the target features and target labels.

For example, stimulus F1 has four features consistent with the prototype F0 and one feature (i.e., antenna) consistent with the prototype J0. The degree of similarity between a test stimulus and the prototype is defined by the number of matching features of the test stimulus to the prototype of the corresponding category (see Tables 1, 2A and 2B).

There were two levels of similarity: high-match and low-match. In the high-match condition, each test stimulus had four features in common with the prototype of the corresponding category and one feature in common with the prototype of the contrasting category. In the low-match condition, each test stimulus had two features in common with the prototype of the corresponding category and three features in common with the prototype of the contrasting category.

Design and procedure

All experiments reported here had a 2 (Test Condition: classification vs. inference) × 2 (Feature Match: high vs. low) within-participants design, and the procedure consisted of two phases: training and testing. The experiments were administered on a 17-inch computer monitor and controlled by E-Prime 2.0 software. Classification and inference test trials were presented in blocks, and the order of the blocks was counterbalanced. The order of test trials within each block was randomized for each participant.

All experiments started with the training phase. At the beginning of training, both adults and children were told that there were two groups of creatures: Flurps and Jalets. They were then presented with creatures (one at a time, with 4000 ms per item), each accompanied by a category label presented auditorily in a carrier phrase (e.g., "This is a Flurp"). The carrier phrase was prerecorded and presented by a computer. The phrase had the same onset as the beginning of the trial, with a total duration of approximately 1800 ms. There were 36 training trails (i.e., 18 Flurps and 18 Jalets), and this part of the experiment lasted for approximately 3 to 4 min. No participant response was required during this phase.

The testing phase, consisting of 92 trials (12 with feedback and 80 without feedback), was administered immediately after the training phase (see Fig. 2 for examples of testing trials). Half of these trials were classification and half were inference, with each trial presented in a self-paced manner. The classification and inference testing conditions differed in what participants needed to predict. On classification trials, participants predicted the label of an item given information about all five features (they were instructed that they would be presented with a creature and would need to decide whether the creature was a Flurp or a Jalet). On inference trials, participants predicted a missing (i.e., covered) feature given the other four features and the label (they were instructed that they would be presented with a creature with a covered body part and would need to decide which body part was under the cover). For both children and adults, this part of the experiment lasted approximately 14 to 15 min.

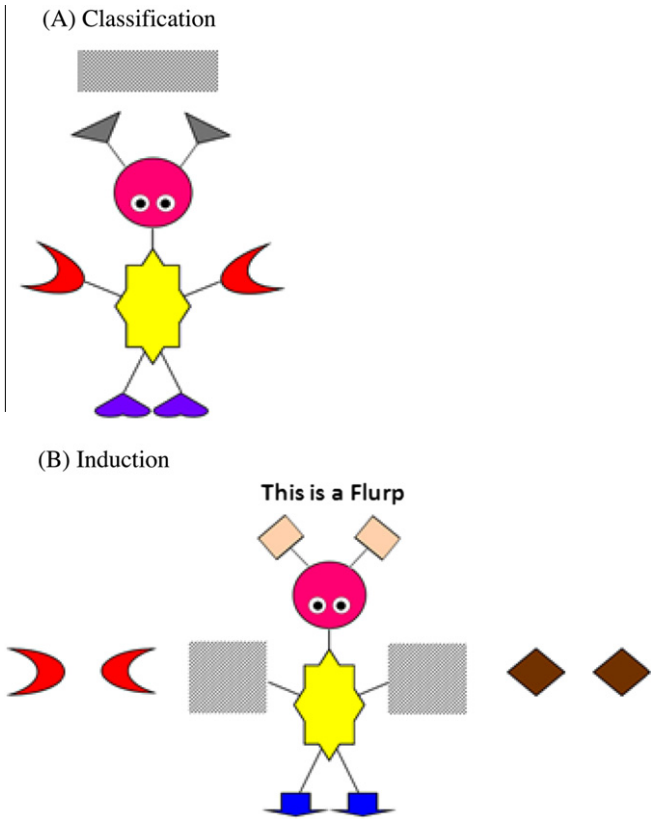

Fig. 2. Examples of classification and induction test trials in Experiments 1 to 4. (A) On classification trials, participants were presented with stimuli and asked whether the item was Flurp or Jalet. (B) On induction trials, participants were presented with stimuli and asked which body part was under the cover.

The procedures were identical for both adult and child participants except for the way the instructions and test questions were presented and the data were recorded. Adults read instructions and the test questions on the computer screen and responded by pressing an appropriate key on the keyboard. For children, all instructions and questions were presented by a female experimenter, and she recorded children's verbal responses by pressing an appropriate key on the keyboard.

To familiarize participants with the testing task, yes/no feedback was given on 12 test trials—the first 6 test trials of classification and induction testing conditions (each of these represented high-match trials). In this and other experiments reported here, children were above 85.2% accuracy on these trials, and adults were above 68.5%, all above chance ($ps < .05$). No feedback was given on the remaining 80 testing trials (40 in each testing condition, half high-match and half low-match), and only these trials were used in the reported analyses. The proportion of responses consistent with the category from which the exemplar was derived (called "category accordance responses" by Yamauchi & Markman, 2000) was the dependent variable.

In addition, a memory check was administered after the main experiment to examine whether participants remembered two categories after completing all of the tasks. There were five memory check trials, with participants being presented with stimuli randomly generated from the training structure (see Table 1). On each trial, participants were asked to recall the corresponding label of each stimulus. Children and adults exhibited memory accuracy of 91.7% and 78.0%, respectively. One adult answered fewer than three of five memory check questions correctly, and these data were excluded from the analysis.

Results and discussion

The main results are presented in Fig. 3. As can be seen in the figure, adults exhibited equivalent performance in the high-match condition (i.e., no differences between classification and inference), whereas there was a marked difference in the low-match condition. In particular, adults were more likely to produce category-consistent responding in the low-match inference condition than in the low-match classification condition. In contrast, young children produced high levels of category-consistent responding only in the high-match conditions, whereas this was not the case in the low-match conditions.

Note that all experiments that involved different age groups (Experiments 1–3) revealed a significant three-way interaction (Age × Testing Type × Feature Match), all Fs > 4.1, ps < .06, $\eta_p^2 > .176$. To interpret the interaction, we conducted separate 2 (Testing Type: classification vs. induction) × 2 (Feature Match: high vs. low) within-participants analyses of variance (ANOVAs) for each age level.

For adults, there was a significant testing type by feature match interaction, $F(1,10) = 24.63$, $MSE = 0.32$, $p = .001$, $\eta_p^2 = .711$. A paired-samples t test indicated that in the high-match condition there were no differences between inference and classification, $t(10) = 0.30$, $p = .772$, whereas in the

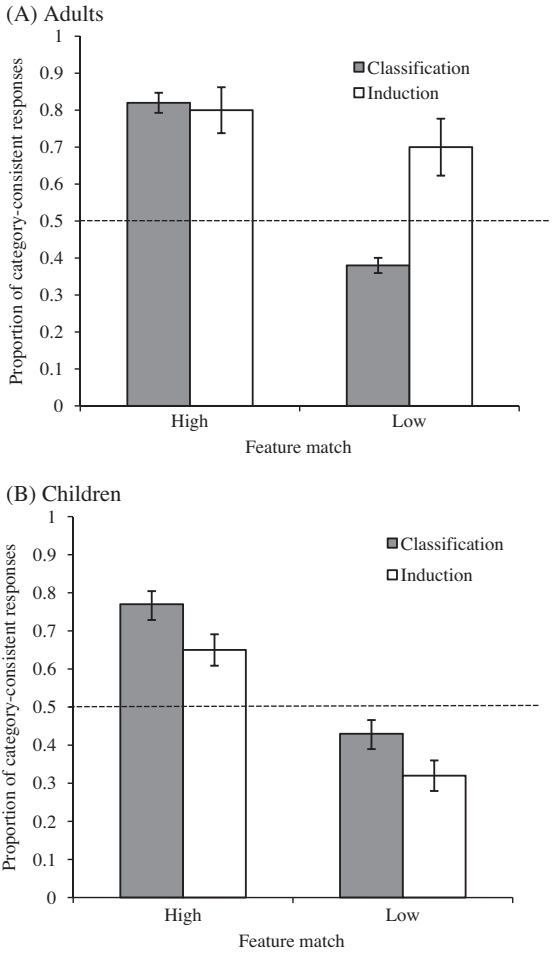

Fig. 3. Proportions of category-consistent responses by feature match and testing condition for adults (A) and children (B) in Experiment 1A. Error bars represent standard error of the mean.

low-match condition participants were more likely to make category-consistent responses in the inference condition than in the classification condition, $t(10) = 3.89$, $p = .003$, $d = 1.71$.

For children, there was a main effect of feature match, $F(1,11) = 43.56$, $MSE = 1.33$, $p = .001$, $\eta_p^2 = .798$, with participants being more likely to provide category-consistent responses in the high-match condition than in the low-match condition. There was also a main effect of testing type (there were more category-consistent responses in the classification condition than in the inference condition), $F(1,11) = 14.77$, $MSE = 0.16$, $p = .003$, $\eta_p^2 = .573$, which was different from adults. This main effect may reflect the advantage of training–testing correspondence (recall that participants in this experiment were trained by classification); we come back to this issue in Experiments 1B and 3.

There were several important differences between children and adults. First, in contrast to adults, for children there was no significant interaction between testing type and feature match, $F(1,11) = 0.02$, $MSE = 0.00$, $p = .90$. Second, unlike adults who were above chance in relying on category information in low-match inference, one-sample $t(10) = 2.66$, $p = .024$, $d = 0.80$, young children performed significantly below chance in relying on the label to predict missing features; they relied instead on the overall similarity, one-sample $t(11) = 4.47$, $p = .001$, $d = 1.29$. Finally, in contrast to adults, children's performance in low-match inference did not exceed that in low-match classification. In fact, the opposite was true; children were somewhat more likely to generate category-consistent responses in low-match classification than in low-match inference, paired-sample $t(11) = 2.32$, $p = .041$, $d = 0.85$.

These results extend those of Yamauchi and Markman (2000), suggesting that labels may play a different role for adults and children. In particular, similar to Yamauchi and Markman's study, adults treated labels differently from other features; in low-match inference, they relied primarily on labels. In contrast, children relied on the overall similarity; the proportion of category-consistent responses in low-match inference was below chance and did not exceed that in low-match classification, providing little evidence that for young children labels are category markers.

In sum, the asymmetry between low-match inference and low-match classification in adults suggests that for adults labels are processed differently from other features. In contrast, children's tendency to rely on the overall similarity in both low-match inference and classification indicates that young children did not treat category labels differently from other features.

Note that Experiment 1A used only classification training, whereas participants were tested in both classification and inference. To ensure that the observed effects are not specific to a particular training condition used in Experiment 1A, we conducted Experiment 1B, in which participants were given inference training.

Experiment 1B

Method

Participants

Participants were 11 adults (6 women and 5 men) and 12 preschool children (mean age = 54.2 - months, range = 50.9–56.9, 4 girls and 8 boys).

Materials, design, and procedure

The materials, design, and procedure were similar to those in Experiment 1A, with several differences. First, in contrast to Experiment 1A, participants were given inference training. Before training, both adult and child participants were told that there were two groups of creatures, with members of each group having something special inside its body. One group of creatures was said to have a Flurp inside its body, whereas the other was said to have a Jalet inside its body. On each training trial, one creature was presented and participants were told, "This one has a Flurp [or a Jalet] inside its body."

There were also some differences in testing. In the classification task, participants were asked to predict the feature label of an item given information about all five features (e.g., "Is it from the group with a Flurp or with a Jalet inside the body?"). In the inference task, they were asked to predict the value of one of five features given the other four features and the label.

Similar to Experiment 1A, a memory check was administered after the main experiment with all child and adult participants, exhibiting memory accuracy of 83.3% and 72.7%, respectively. Two adults answered fewer than three of five memory check questions correctly, and these data were excluded from the analysis.

Results and discussion

The main results are shown in Fig. 4. Patterns of responding were very similar to those in Experiment 1A. In adults, there were no differences between classification and inference in the high-match condition, whereas there was a marked difference in the low-match condition. In particular, adults were more likely to produce category-consistent responding in low-match inference than in low-match classification. In contrast, young children were likely to produce high levels of category-consistent responding in the high-match condition but not in the low-match condition.

The data were submitted to two separate 2 (Testing Type: classification vs. inference) × 2 (Feature Match: high vs. low) within-participants ANOVAs. For adults, there was an interaction, $F(1,8) = 9.92$, $MSE = 0.26$, $p = .014$, $\eta_p^2 = .553$. Specifically, in the low-match condition, participants were more likely

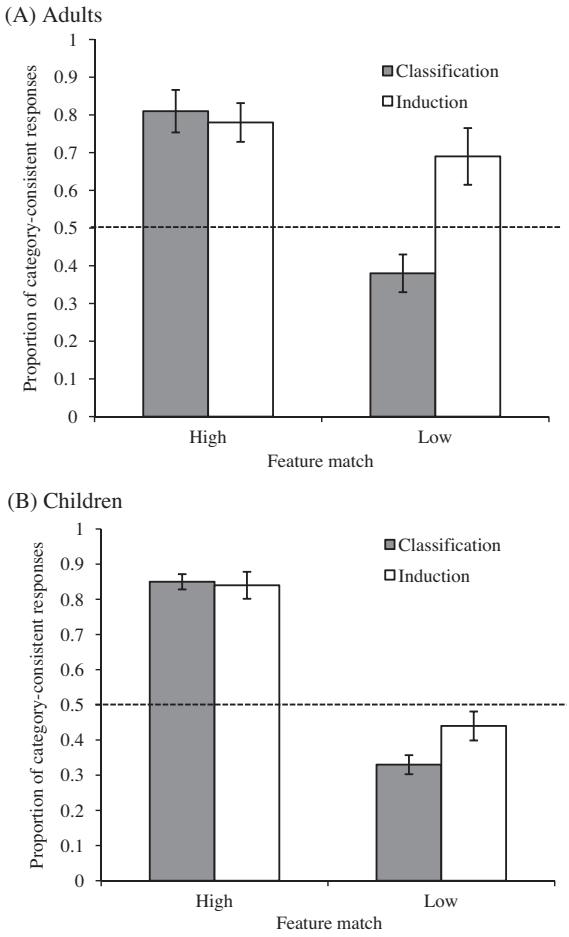

Fig. 4. Proportions of category-consistent responses by feature match and testing condition for adults (A) and children (B) in Experiment 1B. Error bars represent standard error of the mean.

to provide category-consistent responses in inference than in classification, paired-samples $t(8) = 2.79$, $p = .023$, $d = 1.62$, which was not the case for the high-match condition, paired-samples $t(8) = 0.53$, $p = .613$.

For children, there was a main effect of feature match, $F(1,11) = 86.84$, $MSE = 2.50$, $p = .001$, $\eta_p^2 = .888$, with participants being more likely to provide category-consistent responses in the high-match condition than in the low-match condition. There was also a marginally significant interaction, $F(1,11) = 4.22$, $MSE = 0.04$, $p = .064$, $\eta_p^2 = .277$, with participants showing equivalent performance in the high-match conditions, paired-samples $t(11) = 0.11$, $p = .913$, but higher performance in the low-match inference than in the low-match classification, paired-samples $t(11) = 2.79$, $p = .018$, $d = 0.91$. However, unlike adults who were above chance in relying on category information in low-match inference, one-sample $t(8) = 2.57$, $p = .032$, $d = 0.84$, young children were not significantly different from chance ($p = .185$).

Furthermore, a comparison with Experiment 1A (where low-match classification performance was somewhat higher than low-match inference performance) suggests that these differences may indeed be training specific, and we addressed this issue directly in Experiment 3. However, critically, in contrast to adults in Experiments 1A and 1B, low-match inference in young children did not exceed chance performance.

These results extend those of Experiment 1A. Adults again showed asymmetric performance in classification and inference tasks, with performance in low-match induction being above chance. Therefore, when a label indicated one prototype and the majority of perceptual features indicated another prototype, adults tended to rely on the label. In contrast, young children exhibited little evidence of relying on the label. In addition, unlike adults, children exhibited evidence of training-specific effects; we examine these effects further in Experiments 2 and 3.

The goal of Experiment 2 was to examine the generality of effects observed in Experiment 1. In Experiment 1, labels were novel count nouns, and adults, but not children, exhibited evidence of consistent reliance on labels in low-match induction. Would these effects hold for different labels? Would children more readily rely on labels in low-match induction if verbal information is familiar? To answer this question, we conducted Experiment 2A, in which participants were presented with familiar count nouns. To avoid providing children with information they know to be false (e.g., naming the current robot-like stimuli as "bear" or "rabbit"), we used the more general count nouns "friendly pet" and "wild animal."

In Experiment 2B, we further examined the generality of effects by replacing count nouns with descriptors of the habitat (e.g., "lives in the forest" vs. "lives in the sea"). Experiment 2B introduced an important additional control; if adults continue to rely on verbal information, this would indicate that for adults verbal information does not need to be presented in the form of count nouns to be treated as category markers.

Experiments 2A and 2B

Except for the verbal information given to participants, Experiments 2A and 2B were isomorphic to Experiments 1A and 1B (i.e., in Experiment 2A participants were presented with classification training, whereas in Experiment 2B they were presented with induction training). Therefore, we describe these experiments (including relevant procedural information) in Table 3.

The main results of Experiments 2A and 2B are shown in Figs. 5 and 6 and . The overall pattern is strikingly similar to that in Experiment 1. In adults, there were no differences between classification and inference in the high-match condition, whereas there was a marked difference in the low-match condition. In particular, adults were more likely to produce category-consistent responding in low-match inference than in low-match classification. In contrast, young children were likely to produce high levels of category-consistent responding in the high-match condition but not in the low-match condition.

The testing data of Experiment 2A were analyzed with two separate 2 (Testing Type: classification vs. inference) $\times$ 2 (Feature Match: high vs. low) within-participants ANOVAs. For adults, there was a significant interaction between testing type and feature match, $F(1,12) = 36.23$, $MSE = 0.48$, $p = .001$,

Table 3
Overview of Experiments 2A and 2B.

Experiment	Participants	Training details	Testing
2A	13 adults (2 women and 11 men) 9 children (mean age = 56.6 months, range = 52.8–60.1, 5 girls and 4 boys)	Training procedure: Classification Verbal information: "Friendly pet" vs. "wild creature" Memory check: Children (83.3%) Adults (86.2%)	Classification and induction
2B	12 adults (6 women and 6 men) 9 children (mean age = 53.6 months, range = 51.1–58.6, 3 girls and 6 boys)	Training procedure: Induction Verbal information: "This one lives in the forest" vs. "This one lives in the sea" Memory check: Children (84.4%) Adults (76.7%)	Classification and induction

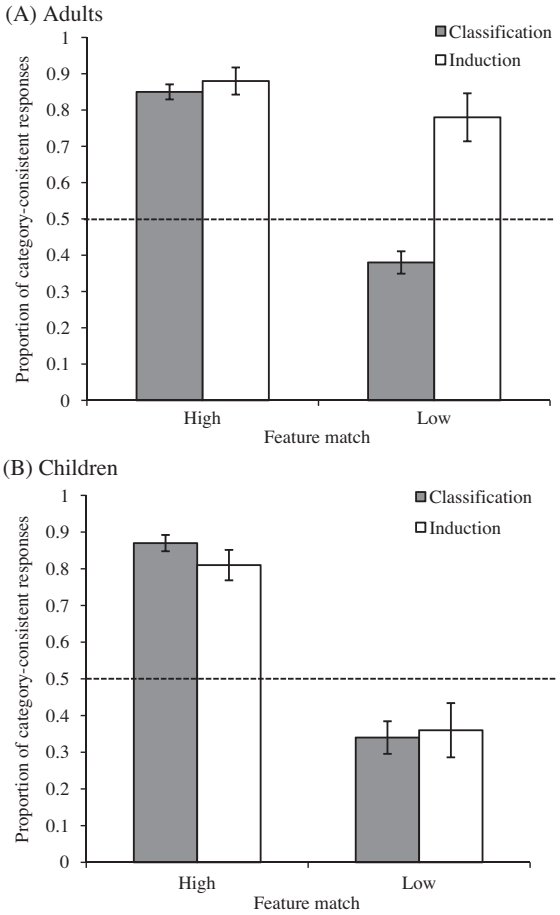

Fig. 5. Proportions of category-consistent responses by feature match and testing condition for adults (A) and children (B) in Experiment 2A. Error bars represent standard error of the mean.

$\eta_p^2 = .751$. Similar to previous results, adults were more likely to provide category-consistent responses in low-match inference than in low-match classification, paired-samples $t(12) = 4.78$,

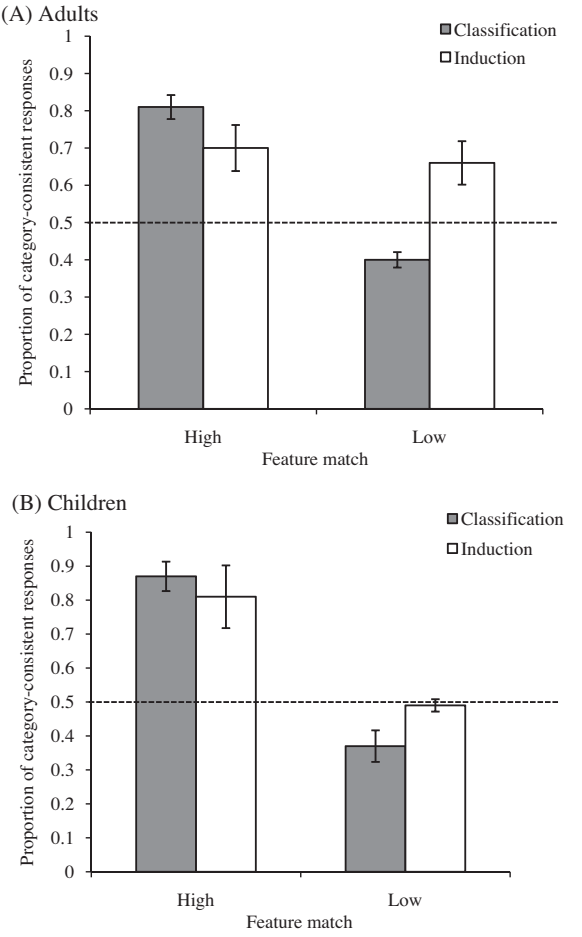

Fig. 6. Proportions of category-consistent responses by feature match and testing condition for adults (A) and children (B) in Experiment 2B. Error bars represent standard error of the mean.

$p = .001$, $d = 2.15$. At the same time, there was no significant difference between the high-match conditions, paired-samples $t(12) = 0.76$, $p = .461$.

For children, there was a significant main effect of feature match, $F(1,8) = 35.66$, $MSE = 2.15$, $p = .001$, $\eta_p^2 = .817$, with children being more likely to provide category-consistent responses in the high-match condition than in the low-match condition. At the same time, neither the main effect of testing nor the interaction approached significance ($ps > .28$). In addition, unlike adults who were above chance in relying on category information in low-match inference, one-sample $t(12) = 4.31$, $p = .001$, $d = 1.19$, young children were marginally below chance, $t(8) = 1.88$, $p = .097$, $d = 0.63$.

Testing data of Experiment 2B were also analyzed with two separate 2 (Testing Type: classification vs. inference) $\times$ 2 (Feature Match: high vs. low) between-participants ANOVAs. For adults, there was a significant testing type by feature match interaction, $F(1,10) = 37.56$, $MSE = 0.38$, $p = .001$, $\eta_p^2 = .790$. Similar to other experiments, adult participants were more likely to give category-consistent responses in low-match inference than in low-match classification, paired-samples $t(10) = 4.08$, $p = .002$, $d = 1.79$, which was not the case for the high-match condition.

For children, similar to the previous experiments, there was a significant main effect of feature match, $F(1,8) = 24.63$, $MSE = 1.50$, $p = .001$, $\eta_p^2 = .755$. Specifically, children were more likely to provide

category-consistent responses in the high-match condition than in the low-match condition. In addition, similar to Experiment 1B, there was a significant testing type by feature match interaction, $F(1,8) = 15.75$, $MSE = 0.07$, $p = .004$, $\eta_p^2 = .663$. Similar to adults, in the low-match condition, they were more likely to provide category-consistent responses in the inference condition than in the classification condition, paired-samples $t(8) = 2.51$, $p = .036$, $d = 1.13$, which was not the case in the high-match condition, paired-samples $t(8) = 0.92$, $p = .38$. However, unlike adults who were above chance in relying on category information in low-match inference, one-sample $t(10) = 2.73$, $p = .021$, $d = 0.83$, young children were not different from chance ($p = .56$).

Overall, results of Experiment 2 replicated and further extended results of Experiment 1. Most critically, differences between children and adults persisted across different ways of presenting verbal information; similar to Experiment 1, adults consistently relied on verbal information, whereas children did not. Furthermore, familiar count nouns (e.g., "friendly pet") used in Experiment 2A did not increase children's reliance on labels, and descriptors (e.g., "lives in the forest") used in Experiment 2B did not attenuate adults' reliance on labels. Therefore, whereas young children exhibited a broad tendency to rely on the overall similarity, regardless of the familiarity or the form of the label, adults tended to rely on labels, also regardless of the familiarity or form of the label.

Finally, a closer examination of the difference between children's performance in classification and induction tasks in Experiments 1A and 2A versus Experiments 1B and 2B suggests that child participants may be affected by different training procedures. In Experiments 1A and 2A, children were trained by classification and during testing they exhibited somewhat better performance in low-match classification than in low-match inference. In contrast, in Experiments 1B and 2B, children were given inference training and during testing they exhibited somewhat better performance in low-match inference than in low-match classification. To equate effects of training, we conducted Experiment 3, in which all participants received both classification and inference training.

Experiment 3

Method

Participants

Participants were 26 adults (8 women and 18 men) and 20 preschool children (mean age = 55.6 - months, range = 48.3–70.0, 13 girls and 7 boys). One child participant was interrupted during the experiment, and these data were excluded from the analysis.

Materials, design, and procedure

The experiment had two between-participants training conditions: (a) classification–label (CL) and inference–descriptor (ID) and (b) classification–descriptor (CD) and inference–label (IL). The orders of CL versus ID and of CD versus IL were counterbalanced. Across the conditions, participants were presented with the same visual stimuli used in previous experiments. Participants were trained with 24 classification trials and 24 inference trials. The corresponding testing trials (46 classification trials and 46 induction trials for each training condition) were administered after all training trials. Similar to previous experiments, yes/no feedback was given on 12 test trials—the first 6 test trials of classification and inference (each of these represents high-match trials). No feedback was given on the remaining 80 testing trials (40 in each testing condition, half high-match and half low-match), and only these trials were used in the reported analyses.

The CL training trials were identical to those of Experiment 1A, and the ID trials were identical to those of Experiment 2B. For example, participants in the CL–ID condition were trained by classification with labels (e.g., "This is a Flurp") and by inference with descriptions (e.g., "This one lives in the forest"). They were then presented with test trials of both classification (to predict the label given all other features) and inference (to predict a missing feature given the other four features and the living place).

The procedure of CD and IL training condition was similar, but the CD trials were identical to those of Experiment 2A and the IL trials were identical to those of Experiment 1B. Specifically, participants in

the CD–IL condition were trained by classification with descriptors (e.g., "This one lives in the forest") and by inference with feature labels (e.g., "This one has a Jalet inside its body"). During testing, participants were presented with test trials of both classification (to predict the descriptor given all other features) and inference (to predict a missing feature given the other four features and the feature label).

Similar to previous experiments, a memory check was administered after the main experiment, and child and adult participants exhibited memory accuracy of 88.4% and 72.8%, respectively. Five adults and one child answered fewer than 6 of 10 memory check questions correctly, and these data were excluded from the analysis.

Results and discussion

Because for both children and adults the effect of training condition (i.e., CL–ID vs. CD–IL) was not significant and did not interact with testing type or feature match (all $ps > .16$), the data were collapsed across two conditions (see Fig. 7).

For adults, there was a significant testing type by feature match interaction, $F(1,20) = 39.09$, $MSE = 0.56$, $p = .001$, $\eta_p^2 = .662$. Specifically, in the low-match condition, participants were more likely to provide category-consistent responses in the inference condition than in the classification condition, paired-samples $t(20) = 4.50$, $p = .001$, $d = 1.26$, which was not the case in the high-match condition, paired-samples $t(20) = 1.79$, $p = .089$.

For children, there was a main effect of feature match, $F(1,17) = 603.60$, $MSE = 2.94$, $p = .001$, $\eta_p^2 = .973$, with participants being more likely to provide category-consistent responses in the high-match condition than in the low-match condition, but in contrast to adults, there was no significant interaction between testing type and feature match ($p > .56$). Furthermore, unlike adults who were above chance in relying on category information in low-match inference, one-sample $t(20) = 2.80$, $p = .011$, $d = 0.61$, young children performed significantly below chance in relying on the label in low-match inference; they relied instead on the overall similarity, one-sample $t(17) = 2.54$, $p = .021$, $d = 0.60$.

Critically, when participants received a combination of classification and inference training, they exhibited the same patterns as when they received only one type of training (Experiments 1 and 2). Specifically, adults relied on labels, whereas children relied on the overall similarity.

To further examine differences between children and adults in their reliance on labels, we analyzed individual patterns of responding of children and adults across all experiments that used count nouns (i.e., Experiments 1, 2A, and 3). Participants who made at least 13 of 20 testing trials category-consistent responses (above chance, binomial $p = .07$) in the high-match inference were selected for the analysis of the response pattern in the low-match induction. Those providing category-based (or label-based) responses on 13 of 20 testing trials were classified as category-based responders, whereas those providing at least 13 of 20 responses based on the overall similarity were classified as feature-based (or similarity-based) responders. The rest were classified as mixed responders. The proportions of label-based, similarity-based, and mixed responders are presented in Table 4. Critically, whereas the majority of adults (i.e., >80%) were consistent label-based responders, only 6% of children were. Instead, children were equally split between similarity-based and mixed responders, and the pattern found in children differed significantly from that found in adults, $\chi^2(1, 69) = 41.3$, $p < .0001$.

Overall, across Experiments 1 to 3 when the label (or descriptor) was pitted against appearance similarity, adults tended to rely on labels when making inductive generalizations, which was not the case for young children. Is it possible that children's failure to rely on labels stemmed from fatigue resulting from multiple trials? Although this possibility seemed unlikely because the overall experiment lasted less than 20 min, we deemed it necessary to compare children's performance for the first and second halves of each experiment. If failure to rely on labels stemmed from fatigue, then reliance on labels should be significantly higher in the first part of the experiment than in the second half. Our analyses of Experiments 1A, 1B, 2A, 2B, 3A, and 3B indicated that this was not the case; in none of the experiments did reliance on labels decrease significantly from the first half of the experiment to the second half (all Bonferroni-adjusted $ps > .70$).

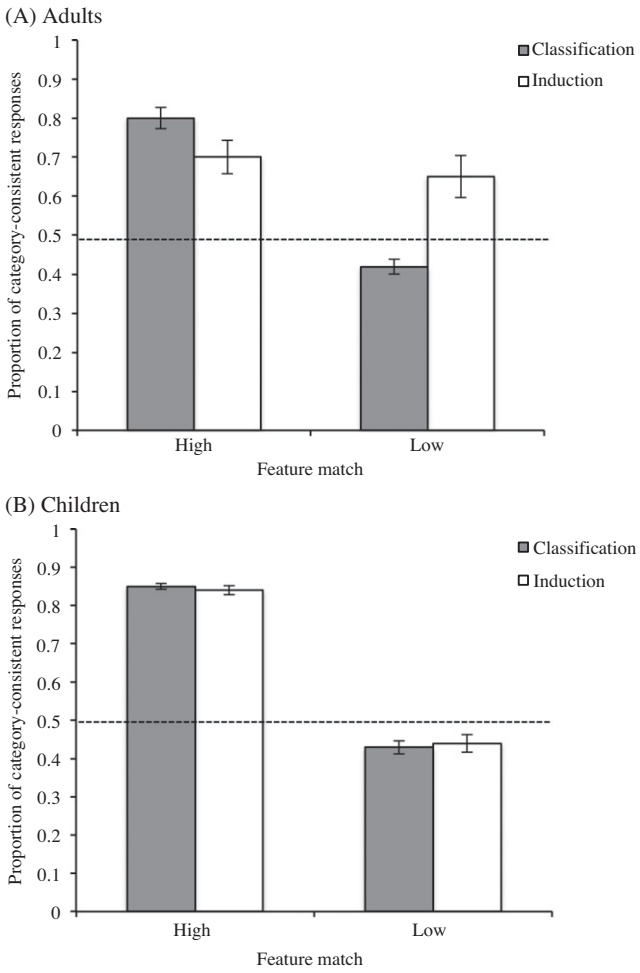

Fig. 7. Proportions of category-consistent responses by feature match and testing condition for adults (A) and children (B) in Experiment 3. Error bars represent standard error of the mean.

Table 4
Numbers (and percentages) of label-based, similarity-based, and mixed responders in Experiments 1, 2A, and 3.

	Children	Adults
Label-based responders	2 (6)	29 (83)
Similarity-based responders	16 (47)	1 (3)
Mixed responders	16 (47)	5 (14)

Note. Percentages are in parentheses.

Taken together, results of Experiments 1 to 3 suggest that whereas for adults labels could be category markers, for young children they are no more than features. Although these results are informative, it could be argued that young children do understand that the labels are category markers, but the results reflect the inability of young children to rely on a single feature when this feature is pitted against multiple features. The goal of Experiment 4 was to address this possibility.

Experiment 4

In Experiments 1 to 3, adult and child participants consistently showed different patterns of responding. The consistent reliance on labels in low-match induction suggests that adults treated labels differently from other features, perhaps as category markers, which was not the case for young children.

In contrast to adults, children's induction was similarity-based; across all of the experiments, they relied on featural overlap rather than on the label (or descriptor) when performing induction. However, whereas Experiments 1 to 3 presented evidence that labels are not category markers for young children, they did not eliminate one important alternative. It is possible that children do understand that the labels are category markers, but they miss the ability to rely on a single feature, especially when this single feature is pitted against multiple features such as in the low-match inference. Although still advancing our understanding of the role of labels in early induction, this latter explanation does not eliminate the possibility that labels are category markers.

To address this issue in Experiment 4, we made one of the nonlinguistic features more salient than any other feature or the label (see Method section for explanation of how this was ascertained). To achieve this goal, the creatures' head was made to move. One type of head motion was consistent with Category 1, and another was consistent with Category 2. The rest of the procedure was similar to that in Experiments 1 to 3.

Method

Participants

Participants were 12 preschool children (mean age = 53.9 months, range = 49.6–59.3, 5 girls and 7 boys).

Materials, design, and procedure

The visual stimuli were identical to those in previous experiments except for the following differences. First, to set up a proper competition between the category information (which did not vary across the exemplars) and a feature, the value of one feature (the head) was also fixed within each category (see Tables 5, 6A and 6B). Second, to make the fixed feature highly salient, the head was animated using Macromedia Flash MX software. The head of one category was pink and moved up and down; whereas for the other category the head was blue and moved sideways. When asked after the experiment what they noticed about the items, all children mentioned the moving head. Two children also mentioned the category label. Therefore, it was concluded that the moving head was more salient than any other feature or the label.

As a result, the learning structure was changed for the part of the head and is shown in Table 5. Similar to previous experiments, Experiment 4 consisted of two phases: training and testing. Two levels of feature match between the test item and the prototype of the corresponding category were

Table 5
Category structure used in learning in Experiment 4.

Category A							Category B						
Stimuli	Body	Hands	Feet	Antenna	Label	Head	Stimuli	Body	Hands	Feet	Antenna	Label	Head
A1	1	1	1	0	1	1	B1	0	0	0	1	0	0
A2	1	1	0	1	1	1	B2	0	0	1	0	0	0
A3	1	0	1	1	1	1	B3	0	1	0	0	0	0
A4	0	1	1	1	1	1	B4	1	0	0	0	0	0
A0	1	1	1	1	1	1	B0	0	0	0	0	0	0

Note. The value 1 = any of six dimensions identical to Category A (see Fig. 1). The value 0 = any of six dimensions identical to Category B (see Fig. 1). A, Category A; B, Category B. A0 and B0 are prototypes of each category, and A1/B1 to A4/B4 are individual exemplars.

Table 6A
Structure of testing stimuli in classification used in Experiment 4.

Category A							Match	Category B						
Stimuli	Body	Hand	Feet	Antenna	Label	Head		Stimuli	Body	Hand	Feet	Antenna	Label	Head
A11	1	1	1	0	?	1	High	B11	0	0	0	1	?	0
A12	1	1	0	1	?	1		B12	0	0	1	0	?	0
A13	1	0	1	1	?	1		B13	0	1	0	0	?	0
A14	0	1	1	1	?	1		B14	1	0	0	0	?	0
A21	0	1	0	0	?	1	Low	B21	1	0	1	1	?	0
A22	1	0	0	0	?	1		B22	0	1	1	1	?	0
A23	0	0	0	1	?	1		B23	1	1	1	0	?	0
A24	0	0	1	0	?	1		B24	1	1	0	1	?	0

Note. High and low are two levels of feature match. A, Category A; B, Category B.

Table 6B
Structure of testing stimuli in induction used in Experiment 4.

Category A							Match	Category B						
Stimuli	Body	Hand	Feet	Antenna	Label	Head		Stimuli	Body	Hand	Feet	Antenna	Label	Head
A11	?	1	1	0	1	1	High	B11	?	0	0	1	0	0
A12	1	?	0	1	1	1		B12	0	?	1	0	0	0
A13	1	0	?	1	1	1		B13	0	1	?	0	0	0
A14	0	1	1	?	1	1		B14	1	0	0	?	0	0
A21	0	?	0	0	0	1	Low	B21	1	?	1	1	1	0
A22	?	0	0	0	0	1		B22	?	1	1	1	1	0
A23	0	0	0	?	0	1		B23	1	1	1	?	1	0
A24	0	0	?	0	0	1		B24	1	1	?	1	1	0

Note. High and low are two levels of feature match. A, Category A; B, Category B.

used: high and low (see Tables 6A and 6B). As shown in Tables 6A and 6B, in the low-match condition there was only one feature (i.e., the moving head) in common with the respective prototype, whereas in the high-match condition there were four such features. The critical condition was low-match inference where only the moving head was in common with the prototype of the corresponding category, whereas three features and the category information (i.e., descriptors of the habitat: "lives in the forest" vs. "lives in the sea") were common with the prototype of the contrasting category. Therefore, if participants rely on multiple features, then they should infer the feature from the contrasting category, thereby exhibiting a high level of category-based responding. In contrast, if they rely on the highly salient moving head, then they should exhibit a low level of category-based responding. In all other conditions, there was no conflict between the category information and the moving head; thus, reliance on the moving head would result in a high level of category-based responding (see Tables 6A and 6B).

The overall procedure was similar to that in Experiment 2B (i.e., the participants received inference training) with the following difference: both the training and testing phases were shortened, with the procedure consisting of 24 training trials and 44 testing trials (similar to previous experiments, the first 12 testing trials were high-match trails accompanied by feedback, and these were not included in the analyses). We shortened the procedure to eliminate the possibility that failure to rely on labels in Experiments 1 to 3 stemmed from fatigue resulting from multiple testing trials. Recall that comparison of the first and second halves of testing in Experiments 1 to 3 undermined this possibility; shortening the procedure allowed us to address this issue directly. Similar to previous experiments, a memory check was administered after the main experiment, with all participants exhibiting high memory accuracy (93%) and with no participant answering fewer than three of five memory check questions correctly.

Results and discussion

The main results of Experiment 4 are shown in Fig. 8. The data were analyzed with a 2 (Testing Type: classification vs. inference) × 2 (Feature Match: high vs. low) within-participants ANOVA. Most important, there was a significant testing type by feature match interaction, $F(1,11) = 129.76$, $MSE = 1.51$, $p = .001$, $\eta_p^2 = .922$. In the high-match condition there was no difference between classification and inference, paired-samples $t(11) = 0.84$, $p = .417$, whereas in the low-match condition participants were more likely to make category-consistent responses in the classification condition than in the inference condition, paired-samples $t(11) = 19.90$, $p = .001$, $d = 8.36$. Most important, when the category descriptor, appearance similarity, and the moving head all indicated the same category (i.e., high-match inference), children were above chance in providing category-consistent responses, one sample $t(11) = 11.86$, $p = .001$, $d = 3.47$. In contrast, when the descriptor denoting a category was pitted against the salient feature (i.e., in low-match inference), children performed significantly above chance in relying on the moving head to infer missing features, one-sample $t(11) = 13.68$, $p = .001$, $d = 3.90$. Therefore, whereas in Experiments 1 to 3 there was no evidence that children rely on label or category information in the low-match induction, they had no difficulty in relying on a single highly salient feature in the current experiment.

Unlike in other experiments reported here, in Experiment 4 children relied on a single feature (i.e., the moving head) rather than on multiple features. Therefore, children's failure to rely on labels in Experiments 1 to 3 is unlikely to stem from their inability to rely on a single feature when it is pitted against multiple features. Although no difference was found between labels and descriptors in Experiments 1B and 2B, it could be argued that results of Experiment 4 would be different had we used count nouns. Therefore, we replicated Experiment 4 with 13 additional children who were given category labels presented as count nouns (e.g., "friendly pet," "wild creature") instead of descriptors. The results of this experiment were equivalent to those of Experiment 4; children were below chance in the low-match inference (31% of category-consistent responses, $t(12) = 8.40$, $p = .001$, $d = 2.33$) and were above chance in the other three conditions (category-consistent responses ranged from 77% to 84%, all $ps = .001$, $ds > 2.60$). Therefore, even when the highly salient moving head was pitted against a count noun, young children relied on the moving head to predict missing features. Furthermore, as shown in a recent study, young children relied on the moving head when labels were presented either as novel count nouns, such as Flurp versus Jalet, or as familiar count nouns, such as carrot-eater versus meat-eater (Deng & Sloutsky, 2012).

Overall, across all of the reported experiments, children failed to rely on either the label or the category descriptor (Experiments 1–3), whereas they relied on a salient perceptual feature

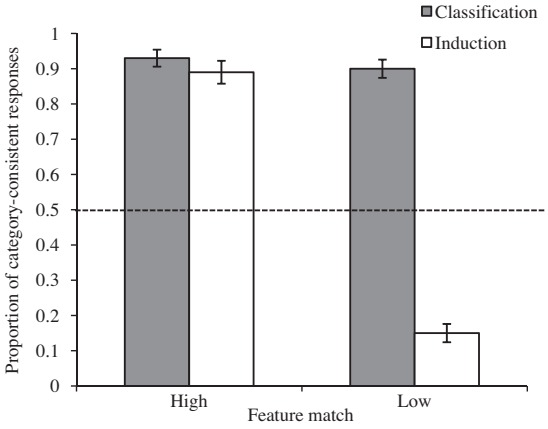

Fig. 8. Proportions of category-consistent responses by feature match and testing condition in Experiment 4. Error bars represent standard error of the mean.

(Experiment 4). In contrast, adults tended to rely on the label (or category information) and not on the overall similarity. These results point to important developmental differences in the role of labels in generalization; whereas adults are likely to treat labels as category markers, there is little evidence that for young children linguistic labels are more than features.

General discussion

The reported research presented six experiments designed to examine the role of labels in early generalization and changes in this role in the course of development. To achieve this goal, we built on the paradigm pioneered by Yamauchi and Markman (1998, 2000). Several major findings stem from the reported experiments.

First, in all experiments, adults relied on category labels when the label was pitted against appearance similarity, which was not the case for young children. In contrast, under no condition did young children exhibit sole reliance on labels in their induction. These effects could not have stemmed from poor category learning or poor memory for labels; across all experiments, children were exceedingly accurate on memory checks, exhibiting memory accuracy of 88%. At the same time, when a highly salient visual feature was introduced (Experiment 4), young children did perform induction by relying on this feature. Taken together, these results offer little evidence that labels are category markers for young children, whereas labels may become category markers in the course of development. As we discuss below, these results have important implication for understanding the development and mechanism of generalization and for theories of categorization.

Labels and the mechanism of generalization

Although researchers agree that from early in development people are capable of performing inductive generalization, the underlying mechanism is a matter of debate. Some have argued that induction is category-based in that when performing induction, people access the category of items in question (see Gelman, 2003, and Murphy, 2002, for reviews). Others have presented an alternative argument that, at least early in development, induction is driven by similarity of compared entities rather than by a common category membership (see Murphy, 2002; Sloutsky, 2010, for reviews). However, under typical circumstances, it is difficult to distinguish between these possibilities. There have been at least two proposals as to how such a distinction could be made.

First, there is an argument that category-based and similarity-based induction may result in different memory traces for studied items, with similarity-based induction resulting in more detailed verbatim memories and category-based induction resulting in less detailed gist-type memory (Fisher & Sloutsky, 2005; Sloutsky & Fisher, 2004b; but see Sloutsky, 2008, and Wilburn & Feeney, 2008). This argument has resulted in a set of studies demonstrating that (a) young children (who presumably perform similarity-based induction) retain more accurate memories of the studied items than adults (who presumably perform category-based induction) and (b) training young children to perform category-based induction attenuates their memory accuracy to the level of adults (Fisher & Sloutsky, 2005; Sloutsky & Fisher, 2004b).

The second idea is to experimentally dissociate category membership and similarity. If induction is category-based, it should follow category information, whereas if it is similarity-based, it should follow similarity information. In one such study (Sloutsky, Kloos, & Fisher, 2007b), 4- and 5-year-olds learned two rule-based categories, with similarity not being predictive of category membership. On learning the rule-based categories, participants were presented with a set of induction trials in which they could rely on either category information or similarity information. Despite the fact that children successfully acquired the categories and retained this knowledge throughout the experiment, their induction was similarity-based (see also Gelman & Waxman, 2007; Sloutsky, Kloos, & Fisher, 2007a, for further discussions; see Griffiths, Hayes, & Newell, 2012, for cases of non-category-based induction in adults).

The third idea is to examine the role of labels in induction; finding evidence that category labels are different from other features and guide inductive inference would support the idea of category-based

induction. The current work found such evidence for adults but not for young children. In addition, if early in development labels are indeed features rather than category markers, then the ability of young children to perform category-based induction is highly questionable. Therefore, current research, in conjunction with previously reported findings (Fisher & Sloutsky, 2005; Sloutsky & Fisher, 2004a; Sloutsky et al., 2001, 2007b), presents further evidence that early induction is similarity-based, but it may become category-based in the course of development.

Language and cognition: are labels features or category markers?

The question of how labels affect generalization is critically important for understanding the mechanism of generalization, but it has broader implications for understanding the role of language in cognition and cognitive development. At the computational level of analysis (Marr, 1982), the labels-as-category-markers approach assumes that words are not merely a part of stimulus input but rather fulfill the role of supervisory signals directing and guiding learning (see Kloos & Sloutsky, 2008 for differences between supervised and unsupervised category learning in children). Thus, if two discriminable items share the same count noun (e.g., both are called a "dax"), the name serves as a top-down signal that the items are equivalent in some way (cf. Gliga, Volein, & Csibra, 2010). In contrast, if words are like any other perceptual feature, then they are part of perceptual input contributing to the overall category structure.

Each of these possibilities presumes a different mechanism and dedicated neural architecture and a different developmental trajectory. Distinguishing among them and understanding the mechanisms underlying the effect of words on category learning is critically important for understanding cognitive development. If from early in development words are supervisory signals, then top-down effects need to play a significant role in early cognitive development. Perhaps the most important implication is that at both the cognitive and neural levels, the lower level processes (e.g., discrimination, generalization) are subject to top-down control. Alternatively, if words become supervisory signals in the course of development, then top-down control does not need to exhibit early onset and could itself be the product of development. Therefore, understanding the role of labels in generalization has implications for most fundamental aspects of cognitive development as well as for understanding the interaction between language and cognition.

Although additional research is needed, the current research indicates that even at 4 or 5 years of age labels function more like features than category markers. When and how do labels become category markers? It seems that there are two possible ways of approaching these questions: pessimistic and optimistic. According to a pessimistic view, the presented results severely undermine the claim that labels do have the special status for young children while not providing conclusive evidence for the special status of the label even in adults. Indeed, results of Experiment 4 suggest that even overwhelming reliance on the label might not be indicative of the fact that the label is a category marker; although children in Experiment 4 overwhelmingly relied on the moving head, we cannot envision a claim that the moving head is a category marker. However, if one assumes a more optimistic view—that labels eventually become category markers for adults—then a theoretical and empirical challenge is to establish the developmental mechanism of this process.

Labels-as-features versus labels-as-category-markers: implications for the relationship between categorization and induction

Recall that much evidence suggests that induction and classification learning are not equivalent for adults, who form different representations in the course of classification and inference training. In particular, under most conditions, categorization training results in the discovery of the features that distinguish among the contrasting categories, whereas inference training results in the discovery of features that are most common in the given category and of inter-feature relations (see Markman & Ross, 2003, for a review; Chin-Parker & Ross, 2004; Sakamoto & Love, 2010; Yamauchi & Markman, 1998; see also Love et al., 2004, for computational modeling).

These differences in representation also manifest themselves in differences in learning rates; for most family resemblance categories, inference training is faster than classification training

(see Markman & Ross, 2003, for a review), whereas the opposite is true for non-linearly separable categories (Love et al., 2004). At the same time, little is known about the development of these differences. Current research reveals no systematic difference between early categorization and induction (sometimes categorization exceeded induction, sometimes the opposite was the case, and sometimes they were statistically equivalent). These findings, in conjunction with evidence that early in development labels function as features, suggest that early categorization and induction could be functionally equivalent. Although we did not examine how children represent categories in the course of classification and inference training (this question is for future research), the current findings allow us to predict that children may form equivalent representations in the course of classification and inference training. If the task does include an exceedingly salient feature (as in Experiment 4), the current results suggest that even in inference training children will learn this highly salient and diagnostic feature rather than the interrelationships among the features (as was the case in previous research with adults). Therefore, the profound differences between classification and inference learning found in adults might not be a fixed property of the tasks; instead, these differences may emerge in the course of development.

Markman and Ross (2003) argued that the differences between categorization and induction pose a challenge to many existing theories of categorization. It seems that the idea that the distinction may emerge in the course of development adds to this challenge.

From features to markers? The changing role of category label in generalization

If we accept that labels do become category markers later in development, it is reasonable to ask what changes in the course of development. One answer can be provided at the computational level; for example, a model of category learning SUSTAIN (Love et al., 2004) introduces a parameter of "category focus" (λ) that governs how much attention is placed on the category label. Depending on the value of the parameter, the label could be similar to other features or could be a category marker. This parameter change offers a mechanistic way of understanding development, but it is also important to understand what triggers this change.

One possible idea that has been discussed elsewhere (e.g., Sloutsky, 2010) is that the contribution of labels to categorization and category learning hinges on (a) the ability to process cross-modal information and (b) the ability to attend selectively. Although neither of these abilities might be sufficient, both seem to be necessary and both may be relatively immature early in development.

First, there is a growing body of evidence that auditory input may affect attention allocated to corresponding visual input (Napolitano & Sloutsky, 2004; Robinson & Sloutsky, 2004; Sloutsky & Napolitano, 2003; Sloutsky & Robinson, 2008). In particular, linguistic labels may strongly interfere with visual processing in prelinguistic infants, but these interference effects may weaken when children start to acquire language (Robinson & Sloutsky, 2008; see also Robinson & Sloutsky, 2007a,b). Given that category learning depends critically on visual processing, labels may hinder learning of new categories in both infants and young children. Therefore, the ability to efficiently process and integrate auditory and visual input appears to be a critical (yet by no means sufficient) step in labels becoming category markers.

Second, for a label to be used as a category marker, participants should be able to selectively attend to relevant information and ignore irrelevant information. However, research published over the past 30 years suggests that young children miss this ability (see Dempster & Corkill, 1999; Hanania & Smith, 2010; Lane & Pearson, 1982, for comprehensive reviews). These difficulties have been linked to the fact that the regions subserving selectivity (most important, the prefrontal cortex) undergo protracted development (Bunge & Zelazo, 2006; Davidson, Amso, Anderson, & Diamond, 2006) and exhibit critical immaturities throughout infancy and the preschool years. In short, the abilities to integrate cross-modal information and to attend selectively seem to be necessary steps for labels to become category markers.

To summarize, the results reported here present evidence that labels may function differently across development; whereas labels are likely to function as features early in development, they may become category markers later in development. Although the abilities to integrate cross-modal information and to attend selectively could be necessary steps in the changing role of labels, precise

mechanisms underlying this transition remain unknown. Therefore, much research is needed to understand why, how, and when labels become category markers.

Conclusion

The current research has presented extensive evidence that (a) early in development labels are features rather than category markers, but they may become category markers in the course of development, and (b) categorization and induction are likely to be equivalent in children but not in adults. The remaining challenge is to understand why and how these transitions take place.

Acknowledgments

This research was supported by a National Science Foundation (NSF) grant (BCS-0720135) and a National Institutes of Health (NIH) grant (R01HD056105) to Vladimir Sloutsky. We thank Catherine Best and Chris Robinson for their helpful comments and Hyungwook Yim for sharing his MatLab script used in simulations.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409–429.
Bunge, S. A., & Zelazo, P. D. (2006). Brain-based account of the development of rule use in childhood. *Current Directions in Psychological Science, 15*, 118–121.
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.
Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia, 44*, 2037–2078.
Dempster, F. N., & Corkill, A. J. (1999). Interference and inhibition in cognition and behavior: Unifying themes for educational psychology. *Educational Psychology Review, 11*, 1–88.
Deng, W., & Sloutsky, V. M. (2012). Carrot-eaters and moving heads: Salient features provide greater support for inductive inference than category labels. *Psychological Science, 23*, 178–186.
Fisher, A. V., & Sloutsky, V. M. (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. *Child Development, 76*, 583–597.
Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. New York: Oxford University Press.
Gelman, S. A., & Heyman, G. D. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological Science, 10*, 489–493.
Gelman, S. A., & Markman, E. (1986). Categories and induction in young children. *Cognition, 23*, 183–209.
Gelman, S. A., & Waxman, S. R. (2007). Looking beyond looks: Comments on Sloutsky, Kloos, and Fisher (2007). *Psychological Science, 18*, 554–555.
Gliga, T., Volein, A., & Csibra, G. (2010). Verbal labels modulate perceptual object processing in one-year-old infants. *Journal of Cognitive Neuroscience, 22*, 2781–2789.
Griffiths, O., Hayes, B. K., & Newell, B. R. (2012). Feature-based versus category-based induction with uncertain categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 576–595.
Hanania, R., & Smith, L. B. (2010). Selective attention and attention switching: Towards a unified developmental approach. *Developmental Science, 13*, 622–635.
Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General, 139*, 319–340.
Kloos, H., & Sloutsky, V. M. (2008). What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General, 137*, 52–72.
Lane, D. M., & Pearson, D. A. (1982). The development of selective attention. *Merrill–Palmer Quarterly, 28*, 317–337.
Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309–332.
Mandler, J., & McDonough, L. (1996). Drinking and driving don't mix: Inductive generalization in infancy. *Cognition, 59*, 307–335.
Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129*, 592–613.
Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.
Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
Napolitano, A. C., & Sloutsky, V. M. (2004). Is a picture worth a thousand words? The flexible nature of modality dominance in young children. *Child Development, 75*, 1850–1870.
Rehder, B., Colner, R. M., & Hoffman, A. B. (2009). Feature inference learning and eyetracking. *Journal of Memory & Language, 60*, 394–419.
Robinson, C. W., & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development, 75*, 1387–1401.
Robinson, C. W., & Sloutsky, V. M. (2007a). Linguistic labels and categorization in infancy: Do labels facilitate or hinder? *Infancy, 11*, 233–253.

Robinson, C. W., & Sloutsky, V. M. (2007b). Visual processing speed: Effects of auditory input on visual processing. *Developmental Science, 10*, 734–740.
Robinson, C. W., & Sloutsky, V. M. (2008). Effects of auditory input in individuation tasks. *Developmental Science, 11*, 869–881.
Sakamoto, Y., & Love, B. C. (2010). Learning and retention through predictive inference and classification. *Journal of Experimental Psychology: Applied, 16*, 361–377.
Sloutsky, V. M. (2008). Recognition memory and mechanisms of induction: Comment on Wilburn and Feeney. *Cognition, 108*, 500–506.
Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive Science, 34*, 1244–1286.
Sloutsky, V. M., & Fisher, A. V. (2004a). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General, 133*, 166–188.
Sloutsky, V. M., & Fisher, A. V. (2004b). When learning and development decrease memory: Evidence against category-based induction. *Psychological Science, 15*, 553–558.
Sloutsky, V. M., & Fisher, A. V. (2008). Attentional learning and flexible induction: How mundane mechanisms give rise to smart behaviors. *Child Development, 79*, 639–651.
Sloutsky, V. M., Kloos, H., & Fisher, A. V. (2007a). What's beyond looks? Reply to Gelman and Waxman. *Psychological Science, 18*, 556–557.
Sloutsky, V. M., Kloos, H., & Fisher, A. V. (2007b). When looks are everything: Appearance similarity versus kind information in early induction. *Psychological Science, 18*, 179–185.
Sloutsky, V. M., & Lo, Y.-F. (1999). How much does a shared name make things similar? 1. Linguistic labels and the development of similarity judgment. *Developmental Psychology, 35*, 1478–1492.
Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels and the development of inductive inference. *Child Development, 72*, 1695–1709.
Sloutsky, V. M., & Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development, 74*, 822–833.
Sloutsky, V. M., & Robinson, C. W. (2008). The role of words and sounds in visual processing: From overshadowing to attentional tuning. *Cognitive Science, 32*, 354–377.
Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12–13-month-old infants. *Cognitive Psychology, 29*, 257–302.
Welder, A. N., & Graham, S. A. (2001). The influences of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development, 72*, 1653–1673.
Wilburn, C., & Feeney, A. (2008). Do development and learning really decrease memory? On similarity and category-based induction in adults and children. *Cognition, 106*, 1451–1464.
Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*, 124–148.
Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 776–795.