How Much Does a Shared Name Make Things Similar? Part 1. Linguistic Labels and the Development of Similarity Judgment

Vladimir M. Sloutsky and Ya-Fen Lo
Ohio State University

Most theoretical proposals considering effects of language on similarity assume that labels affect similarity in a qualitative all-or-nothing manner. This article proposes another theoretical alternative—a model of the label as a discrete attribute of an object. According to this model, the relative weight of labels decreases with the child's age. Predictions derived from the model were tested in 3 experiments. In these experiments, children aged 6–12 years were presented with triads of schematic faces and were asked to make similarity judgments. Similarity of faces within the triads was manipulated via systematic variation of distinct facial features. It was found that (a) labels contribute to similarity judgment in a quantifiable manner, (b) labels' weight decreased with age, and (c) effects of labels are likely to stem from the cross-modality of labels. These results are discussed in relation to theories of development of categorization.

The ability to grasp similarity and to make similarity judgments is ubiquitous in human thought (Estes, 1994; Goldstone, Medin, & Gentner, 1991; Medin, Goldstone, & Gentner, 1993; Tversky, 1977). In particular, the notion of similarity is central to models of (a) categorization- and generalization-based reasoning, such as induction and analogy; (b) memory and memory development; and (c) learning and transfer. For example, models of categorization assume that people group similar objects together and generalize inferences to objects that are similar to the target objects (Lopez, Gelman, Gutheil, & Smith, 1992; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rosch, 1978). Models of memory and memory development assume that similar items are stored together and that the retrieval of an item may activate the retrieval of similar items (Estes, 1994; Ratcliff, 1978; Roediger, 1990). Finally, models of learning and transfer assume that tasks similar to those learned are more likely to be learned and that acquired knowledge is more likely to be transferred to similar items (Brown, 1990; Gick & Holyoak, 1980; Goswami, 1992; Goswami & Brown, 1989; Ross, 1989; Thorndike, 1913). Two questions pertaining to similarity

seem to be critically important. First, what makes distinct objects psychologically similar? Second, how does this psychological similarity develop ontogenetically? Attempts to answer these questions, however, have revealed that the notion of similarity, frequently used in psychological theorizing, is difficult to capture.

In particular, similarity has been construed either as a cognitive *primitive* (Gentner, 1978; Medin et al., 1993; Rosch, 1978; Smith, 1989; Smith & Heise, 1992; Smith, Jones, & Landau, 1996) or as a *product* of domain-specific knowledge (for reviews, see Gelman & Wellman, 1991; Gopnik & Meltzoff, 1997; Keil, 1989; Wellman & Gelman, 1992). According to the first view, the ability to detect similarity does not require prior knowledge, and a small set of low-level, or "dumb," processes (e.g., selective attention to specific perceptual properties) can lead to the development of high-level, or "smart," conceptual knowledge (Smith, 1989; Smith & Heise, 1992). Proponents of this position have argued for causal interrelatedness and interdependence of perceptual and conceptual similarity.

The second view could be illustrated by Goodman's (1972/1972) argument that the similarity between two entities requires one to identify dimensions on which these entities are similar. Therefore, the knowledge of these dimensions and beliefs about their importance should come prior to any similarity judgment. According to the second view, even at the outset of development, children have a repertoire of high-level conceptual knowledge (Gelman & Markman, 1987; Gelman & Wellman, 1991; Keil, 1989; Mandler, 1997; Mandler, Bauer, & McDonough, 1991; Mandler & McDonough, 1996; Soja, Carey, & Spelke, 1991; Wellman & Gelman, 1992). Proponents of this position have distinguished between perceptual (low-level) similarity and conceptual (high-level) similarity and have advocated the relative independence of the two.

In what follows, we consider formal models of similarity and relationships between linguistic labels and similarity. We then outline a quantitative model linking linguistic labels, similarity, and categorization. On the basis of the assumption that linguistic labels are attributes and labels' weights may differ from those of

other attributes, we proposed a low-level mechanism underlying similarity. Finally, we present three experiments designed to test predictions of the model.

Formal Models of Similarity

Most formal models of similarity have considered the ability to grasp similarity to be a cognitive primitive (Estes, 1994; Nosofsky, 1986; Tversky, 1977; but see Medin et al., 1993). Even though the theories differ in many ways,[1] they generally agree that representations in memory are encoded in terms of attributes of objects and events (see Estes, 1994, for a discussion) and that people consider stimuli sharing more attributes (or relations among attributes) to be more similar.[2] These theories further agree that different features and attribute dimensions may have different weights in similarity judgment. As a result, within many formal approaches, similarity has been construed as a weighted combination of shared attributes or attribute dimensions; similarity varies with dimensional weights (see also Smith, 1989, and Smith & Heise, 1992, for discussion). Take Smith and Heise's example: Suppose a participant is presented with a crow, a bat, and a flamingo and is asked to make a similarity comparison by selecting the two most similar entities. In this case, the similarity space could be presented as a two-dimensional space with "overall shape and color" as one dimension and "head and feet features" as another dimension. If attention is focused on the first dimension (overall shape and color weigh more), then the crow and the bat would be judged similar, whereas if attention is focused on the second dimension (head and feet features weigh more), then the crow and the flamingo would be judged more similar. Similarly, Tversky (1977) demonstrated that people deemed Austria to be more similar to Sweden when the choices included Sweden, Poland, and Hungary (they focused on the political distinction between Soviet satellites and Western democracies). However, they deemed Austria to be more similar to Hungary when the choices included Norway, Sweden, and Hungary (in this case, they focused on the geographic distinction between Scandinavian and other European countries). Both examples suggest that similarity judgment may change with a change in dimensional weights. These weights are dynamic (rather than fixed), and they vary with the context (Tversky, 1977), the way of presenting the task (Gentner & Markman, 1997; Kotovsky & Gentner, 1996; A. B. Markman & Gentner, 1997), and the process of development (Kotovsky & Gentner, 1996; Smith, 1989; Smith & Heise, 1992).

Labels and Similarity

One important property that has been linked to similarity judgment is a name or linguistic label of an attribute, object, category, or relation. There are a number of theoretical proposals that consider interrelations between language and similarity judgment (see Gentner & Rattermann, 1991, for a review), including the following: (a) Linguistic labels affect similarity judgment, directing the sampling of attribute dimensions to be used in the judgment and focusing on the relevant dimensions, and (b) similarity among objects affects acquisition of category labels.

The first position entertains the Whorfian hypothesis claiming effects of language on similarity judgments. For example, categorical labels were found to play an important role in young children's recognition of similarity. In particular, when objects were labeled (e.g., "This is a dax. Put it with the other dax"), 2- to 3-year-olds were more likely to group these objects taxonomically (e.g., a police car with another car) than thematically (e.g., a police car with a policeman; E. Markman & Hutchinson, 1984; see also Gentner & Rattermann, 1991, and E. Markman, 1989, for reviews). Gentner and her colleagues (Gentner & Rattermann, 1991; Kotovsky & Gentner, 1996; Lowenstein & Gentner, 1998) have also demonstrated that when higher order relations among attributes have been named (e.g., symmetric stimulus patterns were called "even," or descending size patterns were called "daddy, mommy, and baby"), young children more readily recognized these relations. Other researchers (Balaban & Waxman, 1997; Gelman & Markman, 1987; E. Markman, 1989; Roberts, 1995; Waxman & Markow, 1995; Xu, 1998) have argued that linguistic labels facilitate grouping and categorization in infants and young children.

The second position suggests that salient perceptual similarity can direct generalization of category labels onto novel objects (Gentner, 1978; Landau, Smith, & Jones, 1997; Smith et al., 1996). For example, Smith et al. (1996) presented 3-year-old children and adults with a target set of novel objects. Each object consisted of a distinct base and parts. The objects' functions (e.g., "holds pens") and linguistic labels (e.g., Amlas, Warb, Racol, and Fugle) were also introduced. Then the participants were presented with a set of novel objects that shared either functions or perceptually salient parts with objects in the target set. When asked to name objects in the new set, adults relied both on the objects' functions and on perceptually salient parts, whereas children relied only on perceptually salient parts.

Both positions seem to complement each other theoretically. Indeed, the role of the label and of similarity may change depending on the end point of inference. For example, if the end point is similarity judgment, then the label could affect the recognition of similarity. However, if the end point is naming, then similarity could affect the overextension of the label onto novel objects.

Thus, most researchers agree that linguistic labels and similarity are interrelated. There has also been a general agreement on relationships between labels and other attributes of objects, as labels have been considered to be entities that differ from other attributes and whose effects appear in a qualitative all-or-nothing manner. However, the assumption that young children consider labels to be something qualitatively different from other attributes seems to be too strong. There is nothing logically different between labels and other attributes—both labels and attributes could be helpful in predicting other attributes correlated with them (cf. Anderson, 1990). In fact, it has been shown that infants can use attributes to predict other correlated attributes (Younger & Cohen, 1983, 1986; Younger & Fearing, 1998), whereas young children use labels to predict feature sets correlated with them (Vygotsky,

---

[1] For example, Nosofsky (1986) construed similarity as a symmetric relation, with $Sim(A$ to $B) = Sim(B$ to $A)$, whereas Tversky (1977) construed similarity as an asymmetrical relation, with Sim(North Korea to China) $\neq Sim$(China to North Korea).

[2] These theories, however, do not imply that the environment comprises sets of features; they only claim that enormously rich incoming information is encoded in less rich and more manageable form.

1962). For example, when asked to switch names between *cows* and *dogs*, preschoolers predicted that new *dogs* (i.e., cows) should not have horns, but they should bark.

Assuming that children consider labels as attributes, we hypothesized that at the outset of development, labels weigh more than other attributes but that this weight decreases with age, becoming lower than that of other attributes (reasons for this expected decrease are discussed below). If this theoretical position is plausible, then in the case of younger children, labels should contribute to the child's similarity judgment in a manner similar to color, shape, size, and other discernible attributes. In this case, the label's contribution to similarity judgment should be quantifiable in a manner similar to that of the other attributes and features (see Estes, 1994, Medin, 1975, and Tversky, 1977, for different ways of quantification). Our attempt to formulate a label-as-attribute model by quantifying this contribution is presented in the following section.

Quantitative Model of Label as Attribute

The proposed model (a) assumes that a label is an attribute, (b) considers its weight relative to other attributes, and (c) specifies developmental changes in the weight of the label. Because labeling and similarity are inextricably linked to categorization, we specifically focused on interrelationships among similarity judgment, labeling, and categorization.

Suppose that there are entities i and j that have feature structures F_i and F_j (e.g., has a beak, feathers, and can fly vs. is square, has springs, and can jump), a label L_i (e.g., *bird*), and a familiar category K_{familiar} (e.g., BIRD). Suppose further that this is one's first encounter with the entities. Depending on information about F and L and on task demands, there are several possible directions of inference, including (a) grouping, induction, categorization, and similarity judgment and (b) labeling. For example, if one needs to decide whether these feature structures are similar ($F_i \approx F_j$), one is engaged in similarity judgment. If one needs to decide whether a novel entity belongs to a familiar (or unfamiliar) category ($i \in K_{\text{familiar}}$), one is engaged in categorization. If one needs to decide whether j has a feature given that i has this feature and that their feature structures are similar, one is engaged in induction. In addition, if one needs to decide whether these feature structures should have the same label ($L_i = L_j$) given that they have dissimilar ($F_i \neq F_j$) or similar ($F_i \approx F_j$) feature structures, one is engaged in labeling.

We considered categorization to be (at minimum) a partitioning of a set of n entities in the world into s disjoint sets (Anderson, 1990). These entities have multiple dimensions, and therefore a partitioning can also result in a creation of m hierarchical and collateral sets. Because the total number of possible ways to partition n is intractable as it grows exponentially with n, and because many categories represent clusters of correlated features (Rosch, 1978), it seems unlikely that all different ways of partitioning are equally probable. Although there is no theory specifying these probabilities, there are two earlier mentioned theoretical positions. According to one, categorization may develop from a set of low-level perceptual and attentional processes (Smith, 1989; Smith & Heise, 1992), whereas according to the other, categorization develops from high-level conceptual knowledge (for reviews, see Gelman & Wellman, 1991; Gopnik & Meltzoff, 1997;

Keil, 1989; Wellman & Gelman, 1992). Although both of these alternatives are logically possible, which of these possibilities actually takes place remains an open empirical question. The proposed model of label as attribute uses a small set of low-level mechanisms in an attempt to explain the development of complex processes. Of course, such an attempt is not uncontroversial, but we deemed it worthwhile to try to build explanations without taking much for granted. Our only assumptions were that different attributes and attribute dimensions may have different weights in similarity judgment (cf. Smith, 1989) and that children consider labels to be attributes.

On the basis of these assumptions, we hypothesized that children start out with linguistic labels that have greater weights than most other attributes of objects. It seems that a greater weight of the label, as compared with the other attributes, may have an important function of directing attention to existing similarities (cf. Balaban & Waxman, 1997; Roberts, 1995; Waxman & Markow, 1995). Young infants may need such a mechanism to develop the ability to detect correlations among features—an ability that is absent in 7-month-olds but is present in older infants (Younger & Cohen, 1986).

The greater weight hypothesis can account for several well-established empirical phenomena. However, before we review these phenomena, we introduce for the sake of clarity an important theoretical distinction. It is useful to distinguish between two modes of categorization appearing in the course of development: (a) primarily acquisition of a categorical structure and (b) primarily maintenance and updating of a categorical structure. Although there may be no clear-cut borderline between the two modes, it seems intuitively obvious that there should be a higher base rate of novel categories than the base rate of familiar categories in the former, whereas there should be a higher base rate of familiar categories in the latter. When a child starts acquiring a categorical structure, there are multiple ways of partitioning n; however, when a person becomes committed to a particular categorical structure, it is unlikely that alternative ways of partitioning n will be exercised. A change in the weight of labels may indicate a switch from the acquisition mode to the maintenance mode. A larger weight of labels could steer partitioning in a particular direction, whereas frequent examples of the relative nonimportance of labels could decrease their weights, thus preventing alternative ways of partitioning. We start with the acquisition mode, in which the base rate of novel categories is high.

There are several well-established phenomena indicating that during the acquisition mode, labels might have larger weights than other attributes. It is well-known that if two novel entities share a label and other features, they are likely to be granted a similar, rather than a different, feature structure (E. Markman, 1989; Waxman & Markow, 1995). It has been also established that if two entities have different labels, they are likely to be granted different feature structures, even if they share some features (Gelman & Markman, 1987). These relations are captured by Equations 1 and 2, where P is a probability of F_i *and* F_j being deemed similar or different ($F_i \approx F_j$ or $F_i \neq F_j$) given that they have the same or a different label ($L_i = L_j$ or $L_i \neq L_j$):

$$P(F_i \approx F_j | L_i = L_j) > P(F_i \neq F_j | L_i = L_j) \qquad (1)$$

$$P(F_i \neq F_j | L_i \neq L_j) > P(F_i \approx F_j | L_i \neq L_j). \qquad (2)$$

It has been also demonstrated that younger children (i.e., those for whom the base rate of familiar categories is low and the base rate of novel categories is high), when presented with an object having a novel label, are likely to identify it as a member of a novel category (E. Markman, 1987). Similarly, if a label accompanies an object, the object is more likely to be interpreted as a member of a category than if the object appears without a label (Waxman & Markow, 1995). These relations are depicted in Equations 3 and 4, respectively:

$$P(i \in K_{novel}|L_{novel}) > P(i \in K_{familiar}|L_{novel}) \tag{3}$$

$$P(i \in K|L) > P(i \in K|\neg L). \tag{4}$$

Why does the label's weight decrease with age, and when does the decrease occur? We do not know exactly when such a transition takes place, but we believe that several factors might lead to such a decrease. These include (a) a decrease in the base rate of novel categories; (b) acquisition of homonyms and labels that poorly correlate with other features (e.g., *frequent flyer*); and (c) acquisition of synonyms, collateral, and hierarchical categories. We believe that these factors may serve as counterexamples to the relations denoted by Equations 1–4, thus weakening or degrading those relations. It is well-known that probabilistic relations can be weakened or degraded with an increase in counterexamples that violate those relations (see Rescorla & Wagner, 1972, for a review).

Counterexamples that might degrade relations denoted by Equations 1 and 2 are those that exhibit a dissociation between labels and feature sets. There could be two kinds of dissociation, one where $F \neq F_j|L_i = L_j$, and another where $F_i \approx F_j|L_i \neq L_j$. The former dissociation (i.e., $F_j \neq F_j|L_i = L_j$) includes homonyms— identical labels that refer to completely unrelated objects (e.g., *bat* as an animal vs. as a baseball tool) and labels that poorly predict other attributes (e.g., a New Yorker may be more similar to a relative living in Texas than to another New Yorker[3]). For example, adults may judge objects to be similar even when their labels differ (e.g., the metaphor "my lawyer is a shark"; see Gentner & Markman, 1997, for a review) and may judge other objects to be dissimilar even when they do share labels (e.g., *bat* and *bat*). Findings on homonym acquisition support this notion, indicating that young children have an initial bias against homonyms, though this bias can be overridden by learning (Backscheider & Gelman, 1995). The latter dissociation (i.e., $F_i \approx F_j|L_i \neq L_j$) includes synonyms, or multiple labels that refer to the same object. Because the majority of basic-level categories are also members of subordinate, superordinate, and collateral categories (a doctor could be also a psychiatrist, a woman, and a Southern Baptist), acquisition of subordinate, superordinate, and collateral categories and exposure to synonyms should substantially degrade, or even change, relations denoted by the second inequality. Davidson and Gelman (1990) demonstrated an interesting case of a double dissociation between a label and a feature set. These researchers showed that when a choice set included (a) objects that shared the label but had a different appearance (i.e., $F_i \neq F_j|L_i = L_j$) and (b) objects that had a similar appearance but different labels (i.e., $F_i \approx F_j|L_i \neq L_j$), the probability of label-based choices by 4- and 5-year-olds was quite low.

Finally, with an increase of the base rate of familiar categories, the probability that a novel object that has a novel label will be considered a member of a novel category decreases. Therefore, relations denoted by Equations 3 and 4 could be degraded by an increase of the base rate of familiar categories and the decrease of novel ones.

Thus, we predicted that the weight of labels should decrease (a) with a decrease of novel categories and an increase of familiar categories, (b) in the course of homonymic and synonymic growth, and (c) with the acquisition of subordinate, superordinate, and collateral categories. We could not predict exactly where in the course of development such change takes place, but research on language development and order of acquisition of categories of different levels suggests it takes place somewhere after preschool age (Gelman & O'Reilly, 1988; Sperber, Davies, Merrill, & McCauley, 1982).

Why would labels weigh more for younger children, and what might be a mechanism that underlies the greater weight of labels at earlier age? One such putative mechanism is the cross-modality of labels: Whereas most attributes have a visual input, labels have an auditory input and thus represent a different modality. An attribute that has a different modality from the rest of the attributes may have a privileged processing status and a greater default weight and, as a result, may attract attention to relevant attributes of the "majority" modality. For example, when two objects share a label (name), attention is attracted to the shared visual attributes, whereas visual attributes that are shared with differently labeled objects are not attended to. In fact, the cross-modality hypothesis has been supported empirically. It has been demonstrated that in the case of 15-month-olds, objects were grouped together when they shared a label or a nonlinguistic auditory input (instrumental music input) if either input perfectly correlated with an infant's's fixation of an object (Roberts, 1995; Roberts & Jacob, 1991; but see Balaban & Waxman, 1997). The experiments reported below bring additional empirical support for the cross-modality hypothesis.

In what follows, we consider quantitative aspects of the proposed model. The proposed model of label as attribute has been derived from the product-rule model of similarity (Estes, 1994; Medin, 1975; see also Nosofsky, 1986). According to the product-rule model, similarity among stimuli is a function of the number of overlapping features, and the probability of grouping feature patterns together depends on similarity of these patterns. Similarity of any two patterns i and j each having a total number of N features could be calculated by using Equation 5:

$$Sim(i, j) = S^{N-k}, \tag{5}$$

where N denotes the total number of relevant attributes, k denotes the number of matches, and S $(0 \leq S \leq 1)$ denotes values (weights) of a mismatch. If two entities are identical on all dimensions (there are no mismatches), their similarity should be equal to 1; otherwise, it is smaller than 1.

The product-rule model is based on two important assumptions: (a) The attributes that are relevant (i.e., are attended to) are known, and (b) they all have the same weight. However, each assumption is not necessarily satisfied under all conditions. For example, with respect to the first assumption, there is no formal theory that is

---

[3] We would like to thank an anonymous reviewer for suggesting this example.

Table 1

Estimates of Similarity of the Target (T) to Either Stimulus A or Stimulus B Derived From Equations 5 and 6

Attributes shared by target		No-label condition $Sim(i, j) = S^{N-K}$		Label condition (label shared with B) $Sim(i, j) = S_{label}^{1-L}S_{vis.attr}^{N-k}$			
		Younger children	Older children	Younger children		Older children	
With A	With B	$Sim(T, A)$	$Sim(T, B)$	$Sim(T, A)$	$Sim(T, B)$	$Sim(T, A)$	$Sim(T, B)$
0	0	.125	.125	.031	.125	.090	.125
1	1	.250	.250	.063	.250	.180	.250
2	2	.500	.500	.125	.500	.360	.500
1	0	.250	.125	.063	.125	.180	.125
2	1	.500	.250	.125	.250	.360	.250
2	0	.500	.125	.125	.125	.360	.125

Note. Across conditions, $S_{vis.attr} = .5$. In the no-label condition, S_{label} for younger children $= S_{label}$ for older children. In the label condition, S_{label} for younger children $= .25$; S_{label} for older children $= .71$. $Sim =$ similarity; $S =$ weight of a mismatch; i and $j =$ entities; $N =$ total number of relevant attributes; $k =$ number of matches; $L =$ label match; vis.attr $=$ visual attribute.

capable of predicting which features will be attended to when a stimulus is truly multidimensional. Therefore, we deemed it important to develop unambiguous stimuli with a small number of attributes. With respect to the second assumption and according to aforementioned theoretical considerations, we explicitly contend that the label weighs more than the other attributes. Therefore, we modified Equation 5 so that it would reflect this theoretical position:

$$Sim(i, j) = S_{label}^{1-L}S_{vis.attr}^{N-k}, \tag{6}$$

where N denotes the total number of attributes (including the label), k denotes the number of matches, $S_{vis.attr}$ denotes values (weights) of a mismatch on a visual attribute, S_{label} denotes values of label mismatches, and L denotes a label match. When there is a label match, $L = 1$ and $S_{label} = 1$; when there is a label mismatch, $L = 0$ and $S_{label} < 1$.

According to our theoretical model, for younger children the label weighs more than the other attributes, whereas for older children it weighs less than other attributes, and therefore the label match/mismatch should have stronger effects on similarity judgments of younger children. Overall, it seems reasonable to use Equation 6 to derive approximate estimates of similarity among the stimuli consisting of a small number of distinct attributes. However, to derive such estimates, one must first determine the value of $S_{vis.attr}$. These values could be set a priori or derived from a large data set. However, we could not locate such data sets, so as other researchers (e.g., Estes, 1994) have done, we set the value of $S_{vis.attr}$ equal to .5. Because S_{label} in younger children is predicted to have a larger contribution than $S_{vis.attr}$, whereas S_{label} in older children is predicted to have a smaller contribution than $S_{vis.attr}$, we set the former[4] as $S_{label/younger} = S_{vis.attr}^2 = .25$, whereas we set the latter as $S_{label/older} = S_{vis.attr}^{1/2} = .71$. In the remainder of this section, we describe our experimental approach and specific predictions that are based on Equation 6. Because the parameters have been set a priori, we considered these predictions (although they were expressed quantitatively) to be qualitative ones.

The overall experimental idea used in the reported experiments is as follows. Suppose that there are three stimuli (A, B, and T).

These stimuli have three attribute dimensions—color, shape, and size. Stimulus T is a target, whereas Stimuli A and B are test stimuli; the goal is to decide which of the test stimuli is more similar to the target. Further suppose that Test Stimuli A and B may share no attributes, one attribute (e.g., same size), or two attributes (e.g., same size and color) with the target. Now imagine that A, B, and T are presented under the following two conditions. In the no-label condition, neither stimulus is labeled. In the label condition, stimuli are labeled and, in addition to common attributes, Test Stimulus B also shares a label with the target.

Similarity estimates based on the modified product-rule model for the label and no-label comparisons of the described object set are presented in Table 1. Given the proposed hypothesis, we assumed that in the no-label condition, the number of attribute dimensions was limited to three (e.g., color, shape, and size), whereas in the label condition, the number of attributes equaled four (e.g., color, shape, size, and label). Similarity estimates for the nonlabel condition were based on Equation 5, whereas they were based on Equation 6 for the label condition. Recall that for the younger children, $S_{label} = .25$, whereas for the older children, $S_{label} = .71$; in both groups for all other attributes $S_{vis.attr} = .5$. Recall further that in the label condition, Test Stimulus B always shared the label with the target. To test the hypothesis that the label contributes more to similarity judgment than do other attributes, we deemed it necessary to pit labels against other attributes. We therefore included in the table only those attribute patterns in which Test Stimulus B shared with the target no more visual attributes than did Test Stimulus A.

To estimate the probability of selecting Stimulus B (or Stimulus A) as more similar to the target, we used Equations 7 and 8:

$$P(B) = \frac{Sim(T, B)}{Sim(T, B) + Sim(T, A)} \tag{7}$$

$$P(A) = 1 - P(B). \tag{8}$$

---

[4] Because .25 and .71 are both smaller than 1, their impact is an inverse of their value. Therefore, the impact of .71 is smaller than the impact of .25.

The equations seem to meet intuitive expectations regarding similarity judgment. For example, when both test stimuli are equally similar to the target $(Sim(T, B) = Sim(T, A))$, then $P(B) = P(A)$, with both equal to chance. At the same time an increase in similarity between one of the test stimuli and the target and a simultaneous decrease in similarity between the second test stimulus and the target lead to an increase in the probability of choosing the first stimulus.

Predicted probabilities of choosing Test Stimulus B were derived from Equation 7. These probabilities are presented in Table 2. These predictions were tested in the reported experiment. Because values of S were set a priori, we considered predictions presented in Table 2 to be qualitative (e.g., chance vs. above chance vs. below chance) rather than precise quantitative predictions.

In addition to the cross-modality hypothesis formulated above and predictions presented in the table's cells, we made the following critical predictions:

1. In the no-label condition, there should be no differences in similarity judgments of younger and older children (columns 3 and 4 of Table 2).

2. In the label condition, younger children should be more likely than older children to consider a test stimulus that shares the label with the target to be more similar to the target (columns 5 and 6 of Table 2).

3. Because the label represents an attribute, it should contribute to the similarity judgment in a quantitative manner rather than in an all-or-nothing manner (as might be suggested by variants of the Whorfian hypothesis). In particular, when both test stimuli share some attributes with the target, and Test Stimulus B shares the label, the probability of selecting B over A will vary with the number of attributes shared by A and B with the target and with the age of the child.

Experiment 1

Method

Participants. A total of 107 children aged 6 to 12 years participated in the study. The participants represented three age groups: 34 five-to-seven-year-olds ($M = 6.3$ years, $SD = 0.39$; 14 boys and 20 girls), 41 seven-to-nine-year-olds ($M = 7.8$ years, $SD = 0.4$; 21 boys and 20 girls), and 32 nine-to-eleven-year-olds ($M = 9.3$ years, $SD = 0.9$; 20 boys and 12 girls). The participants were recruited from elementary and middle schools located in middle-class suburbs of Columbus, Ohio. The majority of participants were Caucasians.

Materials. The design included an experimental (label) and a control (no-label) condition. The conditions varied across participants. In both conditions, participants were presented with triads of 2 in. × 2 in. (5.1 cm × 5.1 cm) schematic faces, two of which were test stimuli and one of which was a target. The participants had to select which of the test stimuli faces was more similar to the target. Each schematic face had three distinct attributes (shape of head, shape of ears, and shape of nose), and each attribute had three values (e.g., curve-lined nose, straight-lined nose, and angled nose). A target stimulus could share zero, one, or two attribute values with each of the test stimuli. In the experimental (label) condition, both test stimuli (Test Stimuli A and B) shared attributes with the target, whereas it was always one test stimulus (Test Stimulus B) that shared the linguistic label (an artificial word) with the target. No labels were introduced in the control condition. The design also included six within-subject stimulus-pattern conditions: (a) Pattern T-0-0—the target stimulus shared zero attributes with each of the test stimuli. (b) Pattern T-1-0—the target shared one attribute with Test Stimulus A. (c) Pattern T-1-1—the target shared one attribute with each of the test stimuli. (d) Pattern T-2-1—the target shared two attributes with Test Stimulus A and one attribute with Test Stimulus B. (e) Pattern T-2-0—the target shared two attributes with Test Stimulus A. (f) Pattern T-2-2—the target shared two attributes with both test stimuli. In each of the six stimulus-pattern conditions, participants had three trials. Examples of the stimuli in the condition T-1-1 are presented in Figure 1.

Smith et al. (1996) demonstrated that the overall shape (in this case the shape of the face) and added details (in this case the nose and ears) may have different effects on similarity judgment. Therefore, in the present experiment, these attributes were varied systematically within each of the six stimulus-pattern conditions.

Procedure. A female researcher interviewed children in a quiet room in their schools. Before the experimental task, children were introduced to some warm-up questions and were given feedback. In the warm-up tasks, children were presented test and target stimuli and were asked to choose the test stimulus that was more similar to the target.

In the first warm-up trial, participants were presented with a target (a red triangle) and two test stimuli (a blue triangle and a yellow hexagon). In the second warm-up trial, they were presented with a yellow hexagon as the

Table 2

Predicted Probability of Similarity Judgments Between Stimuli T and B
With Different Weights of the Label

Attributes shared by target		Condition			
		No shared label		Label shared with B	
With A	With B	Older children	Younger children	Older children	Younger children
0	0	.5[a]	.5[a]	.58[a]	.8[b]
1	1	.5[a]	.5[a]	.58[a]	.8[b]
2	2	.5[a]	.5[a]	.58[a]	.8[b]
1	0	.33[c]	.33[c]	.4[a]	.65[b]
2	1	.33[c]	.33[c]	.4[c]	.53[a]
2	0	.2[c]	.2[c]	.26[c]	.5[a]

Note. In label and no-label conditions, probabilities were calculated according to Equation 7. For older children, the label weighed less than the other attributes; for younger children, the label weighed more.
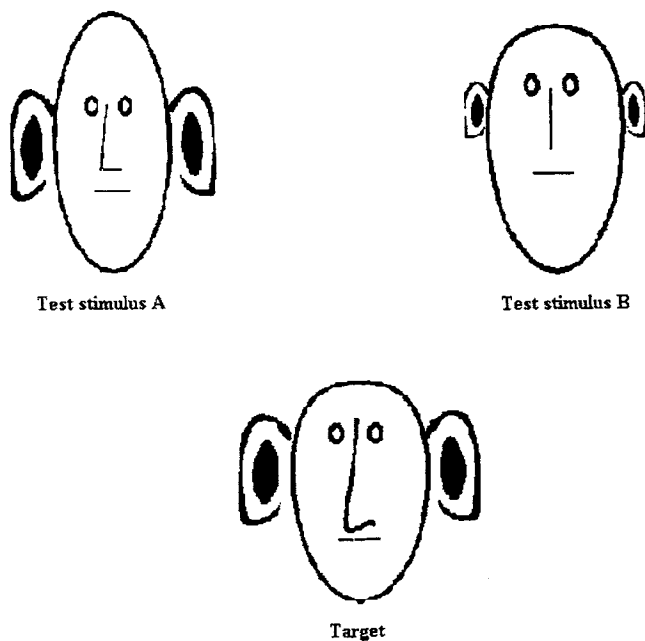[a] Chance. [b] Above chance. [c] Below chance.

Figure 1. Stimuli presented in one trial in the T-1-1 condition: The target (T) shared the shape of the face with Test Stimulus B and the size of ears with Test Stimulus A.

target and a yellow rectangle and a blue triangle as the test stimuli. In the third warm-up trial, children were presented with three pictures of two birds and one turtle. One crane with a round shape was the target; a goose with wings stretched and the turtle with a round shape were the test pictures. In all these warm-up trials, children were asked to determine which of the two test stimuli was more similar to the target. When children failed to answer the similarity questions, the researcher explained how each of the test stimuli could be grouped with the target.

If the child was capable of grouping objects in two out of three warm-up trials, the researcher proceeded to the main experiment (no child was eliminated from the study because all participants grouped objects correctly in at least two out of three warm-up trials). As in the warm-up tasks, in the experimental trials children were also asked which one of the test pictures was more similar to the target. Positions of two test pictures were counterbalanced across the experimental trials. In both the label and no-label conditions, participants had 18 experimental trials (six within-subject stimulus patterns with 3 trials each).

Stimuli for each of the six stimulus-pattern conditions were placed in six separated boxes. Within each box there were five sets of pictures. In each of the 18 trials, the researcher haphazardly selected the box and the set within each box. Because there were five sets of pictures for three trials, only three out of the five sets in each box were chosen.

In both the label and no-label conditions, stimuli were introduced as pictures of aliens that come from different planets. In the label condition, pictures were given an artificial name (e.g., Guga, Bala), and children were asked to repeat the names. If a child failed to do so, the researcher repeated the names again. Then children were presented with the sets of pictures and asked to select one test stimulus that was most similar to the target. After the answers were provided, children were asked to explain their choices. The following passage was the important part of the instructions:

I am going to show you some pictures of aliens so you'll learn more about them. Are you ready to start? Let's start! Here we have three alien pictures [pictures were introduced at this point]. They come from different planets (e.g., Guga and Bala). Could you please repeat these

names? Look at this one. This is a Guga [points to the target]. This is a Bala [points to Test Stimulus A], and this is a Guga [points to Test Stimulus B]. Is this Guga [points to Test Stimulus B] more similar to this Guga [points to the target], or is this Bala [points to Test Stimulus A] more similar to this Guga [points to the target]?

The order of introduction of the test stimuli, their location relative to the target, and the order of asking similarity questions were randomized.

Results and Discussion

In this section, we consider children's similarity judgments, consistency of these judgments across trials, and children's explanations of their judgments.

Figure 2 presents the percentages of responses selecting Test Stimulus B as more similar to the target by age group, labeling condition (label vs. no label), and the stimulus-pattern condition. Several aspects of the figure deserve special consideration. First, data in Figure 2 are mostly consistent with the qualitative predictions presented in Table 2. As predicted, in the youngest group, the label-based similarity judgment was at the above-chance level for conditions T-0-0 ($p < .01$, confidence interval [CI] = 56–87%), T-1-1 ($p < .05$, CI = 51–75%), and T-2-2 ($p < .05$, CI = 54–79%). At the same time, the label-based similarity judgment was at the chance level for conditions T-1-0, T-2-1, and T-2-0. Also as predicted, in the oldest group, across all six conditions, the label-based similarity judgment was at or below chance. In particular, it was at the chance level in conditions T-0-0, T-1-1, T-2-2, and T-1-0, whereas it was below chance in conditions T-2-1 and T-2-0 ($ps < .01$, CIs = 41–37%, respectively).

Second, these data point to differences between the label and no-label conditions in both younger groups but not in the older group. The significance of these results was established by aggregating the number of B choices across the six stimulus-pattern conditions and subjecting them to a 3 (age group) × 2 (labeling condition) analysis of variance, followed by post hoc Tukey tests and planned contrasts. The results supported our predictions: There
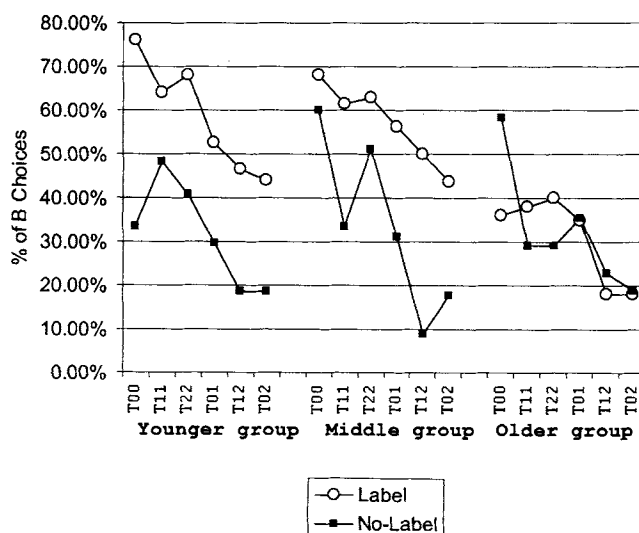

Figure 2. Percentages of Test Stimulus B choices by stimulus pattern condition, age group, and labeling condition. T = target.

was a main effect of age group, $F(2, 101) = 3.9, p < .05$, a main effect of labeling condition, $F(2, 101) = 9.7, p < .01$, and an interaction of age group and labeling condition, $F(2, 101) = 3.7, p < .05$. As predicted, for both groups of younger children there were significant differences in the number of B choices between the label and no-label conditions (for the youngest age group, $M = 10.5$ in the label condition vs. 5.7 in the no-label condition; for the intermediate age group, $M = 10.2$ in the label condition vs. 6.1 in the no-label condition, $ps < .05$). There were no such differences for the oldest group ($M = 5.2$ in the label condition vs. 5.8 in the no-label condition, $p = .8$). Also, as predicted (see Table 2), across the six stimulus-pattern conditions, the number of B choices decreased with age. Although there were no differences between the youngest and the intermediate groups ($M = 8.1$ for each group, $p = .96$), there were significant differences between each of these groups and the oldest group ($M = 5.5$ for the oldest group, $ps < .03$). Finally, as predicted, there were no systematic differences between similarity judgments of younger and older children in the no-label condition ($M = 5.7$ for the youngest, 6.0 for the intermediate, and 5.8 for the oldest group).

The label-as-attribute model based on the product-rule approach to similarity (Equations 6 and 7) also yields a reasonably good quantitative fit between theoretical and observed probabilities of B choices (see Figure 3). Each data point in Figure 3 represents a cell in Table 2, and predicted probabilities correspond to those in the table. In addition to a high correlation between the predicted and observed probabilities ($r = .87$), the proposed theoretical model accounts for approximately 75% of the observed variance ($R^2 = .76$). These findings support the notion that labels contribute to the similarity judgment in a quantitative manner and that this contribution varies with the number of attributes shared by A and B with the target and with the age of the child.

However, it could be argued that these data do not rule out an alternative possibility that within individuals, labels contribute to similarity judgment in a qualitative all-or-nothing manner, whereas the presence of several patterns of responses in the sample makes results look as if responses vary quantitatively. To examine this possibility, we conducted an analysis of individual patterns of responses. Our reasoning was as follows. Within each stimulus pattern, each child was presented with three trials. Therefore, three response types were possible: (a) label-consistent responses (label-

based responses in all three trials), (b) physical-consistent responses (responses based on physical similarity in all three trials), and (c) inconsistent (a combination of label-based and physical-similarity-based responses). If there was indeed a bimodal pattern of qualitative responses, then (a) response types should not vary across stimulus-pattern conditions and age groups, and (b) responses should not vary across trials, nor should there be any inconsistent responses, or else only a small number of them. Proportions of response types in the label condition across the six stimulus-pattern conditions are presented in Figure 4.

Qualitatively, data in Figure 4 indicate that all response types, including inconsistent responses, were present and that proportions of response types varied with stimulus-pattern condition and age. More specifically, in those conditions in which both test stimuli shared equal numbers of attributes with the target (the top three graphs in Figure 4), there were significantly more children in the two younger groups who were label-consistent than who were physical-consistent, $\chi^2(1, N = 298) = 12.3, p < .0001$. At the same time, in those conditions in which Test Stimulus A shared more physical attributes with the target than with Test Stimulus B, whereas B shared the label (the bottom three graphs in Figure 4), this difference was not significant, $\chi^2(1, N = 150) = 0.18, p = .73$. These data, in conjunction with the data on similarity judgment, support critical predictions that in younger children, labels contribute quantitatively rather than in an all-or-nothing manner to the similarity judgment.

Recall that after making a similarity judgment, participants were asked to explain their choices ("Why do you think that this one is more similar to that one?"). Three categories of responses appeared to be most frequent: (a) reference to the label (e.g., "because this is a Bala and this is a Bala"), (b) reference to physical attributes, and (c) reference to both labels and attributes. Proportions of these categories of responses in the label condition by age and stimulus pattern are presented in Figure 5. These proportions were aggregated across the six conditions and subjected to Mann-Whitney tests, with age group as a variable. The analyses indicated that the proportion of label-based explanations was significantly higher in the youngest group than in the oldest group, Mann-Whitney $U(1) = 292, p < .01$. These findings corroborated the results reported above on similarity judgment and consistency of this judgment. Recall that the contribution of labels to similarity judgment and the proportion of label-consistent responses also decreased with age.

In short, the results of the experiment indicated that labels make an important and quantifiable contribution to young children's similarity judgment, but their importance markedly decreases with age. Second, proportions of selecting Test Stimulus B by younger children in the label and no-label conditions differed significantly, whereas there were no such differences for the oldest group. Finally, there were no systematic differences in proportions of selecting Test Stimulus B by younger and older children in the no-label condition.

However, these data do not rule out several alternative explanations. First, the same effects could have stemmed from another possibility—an inability or unwillingness of young children to ignore task-irrelevant information. Indeed, it might be easier for older children to ignore task-irrelevant information than for younger children to do so. If one assumes that younger and older children could deem labels to be task irrelevant (which is not an
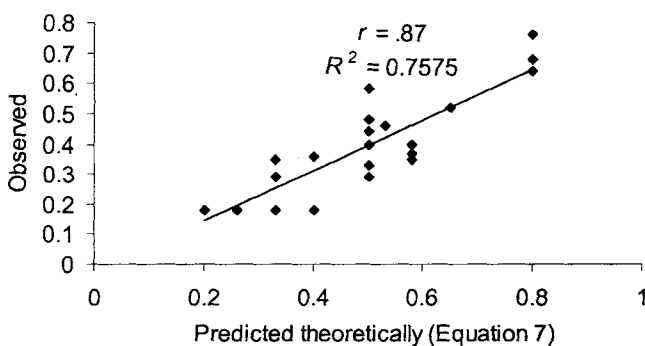

Figure 3. Theoretical probabilities (computed from the model by means of Equation 7) and observed probabilities of Test Stimulus B choices. Each data point represents a cell in Table 2.
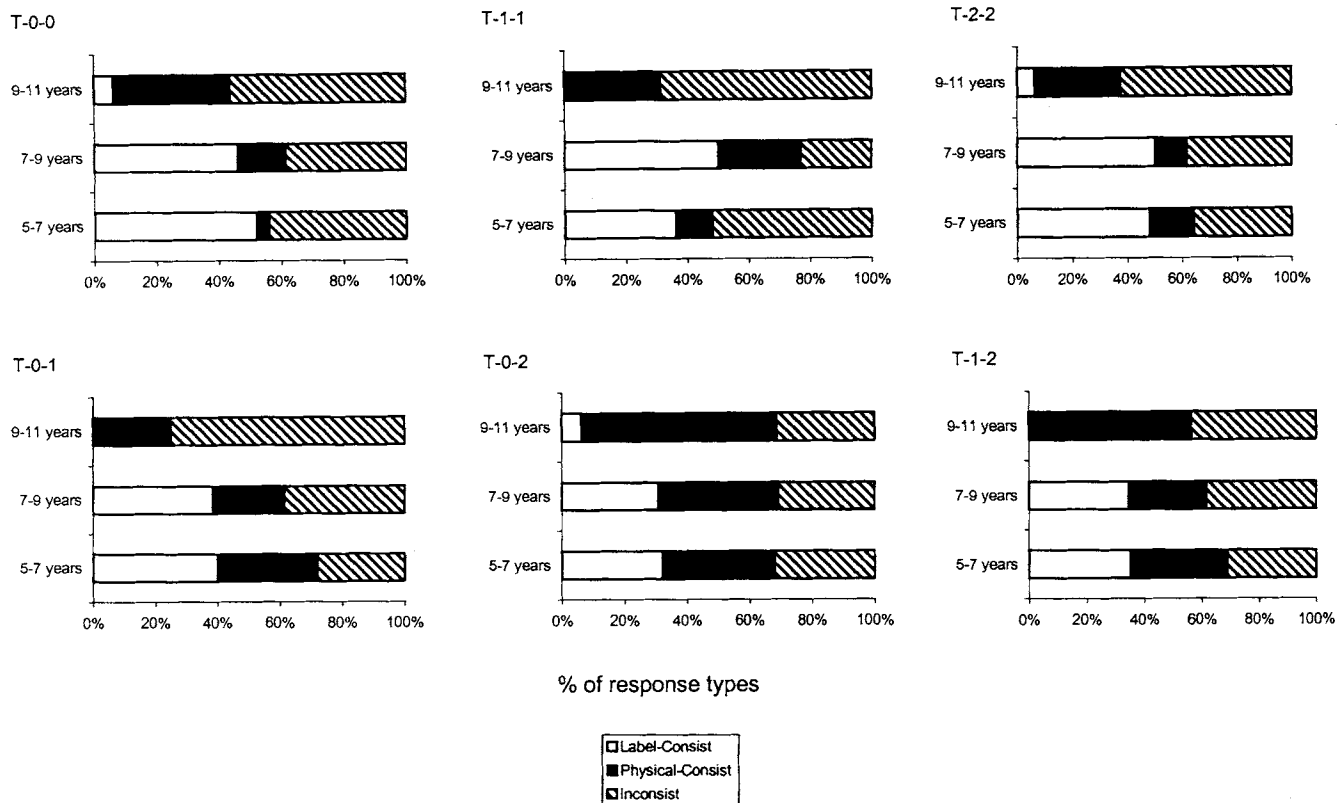
Figure 4. Percentages of label-consistent and physical-consistent responses in the label condition by age and pattern-stimulus condition. T = target; Consist = consistent; inconsist = inconsistent.

unreasonable assumption), then there is little surprise that labels have a stronger effect on the similarity judgment of younger children. It is also possible that younger children merely considered labels to be "tiebreakers," whose effects could be due to younger children's lack of confidence[5] when visual attributes did not allow them to make firm judgments (e.g., in the T-0-0, T-1-1, and T-2-2 conditions). If these arguments are true and the observed effects stem from young children's inability to ignore task-irrelevant information or from their lack of confidence, then their response to other task-irrelevant features should be similar to their responses to labels. Furthermore, these task-irrelevant features should also serve as tiebreakers when visual attributes do not allow children to make a firm judgment.

Second, the cross-modality hypothesis requires additional evidence. For example, it is possible that the reported effects of labels stem not from their being presented in a different modality but from their representing the language (cf. Namy & Waxman, 1998). If this argument is true, then those labels that have linguistic properties (i.e., they denote objects) but that are presented without sound (e.g., signs of a sign language) should have effects similar to those of labels. However, given children's age and familiarity with sign languages, should effects of signs be different from those of labels, there would be strong support for the cross-modality hypothesis.

To test these alternative possibilities, we designed Experiments 2 and 3. In Experiment 2, we presented children with the same pictures as in Experiment 1 except that colored dots (task-

irrelevant features) were substituted for labels. In Experiment 3, we presented children with the same pictures as in Experiments 1 and 2; however, signs representing a sign language were substituted for labels.

Experiment 2

Method

Participants. A group of 14 children aged 6 to 7 years participated in the study ($M = 7$ years; $SD = 0.6$; 8 boys and 6 girls). The participants were recruited from an elementary middle school located in middle-class suburbs of Columbus, Ohio. All the participants were Caucasians.

Materials. The experiment had the same materials as those used in Experiment 1 except that in the reported experiment, labels were substituted with colored dots. These dots were referred to as "the color of alien spaceships." As in Experiment 1, in T-0-1, T-0-2, and T-1-2 conditions, the test stimulus that had fewer overlapping features with the target shared the color of the dot. In the T-0-0, T-1-1, and T-2-2 conditions, it was a randomly selected test stimulus that shared the color of the dot with the target. There was a difference in the design of Experiment 2 as compared with that of Experiment 1—we did not include a control condition in this experiment.

Procedure. The procedure was identical to that of Experiment 1. The only difference was that in the reported experiment, labels were substituted with colored dots. These dots were referred to as "the color of alien

---

[5] We thank James L. Dannemiller for pointing to this possibility.
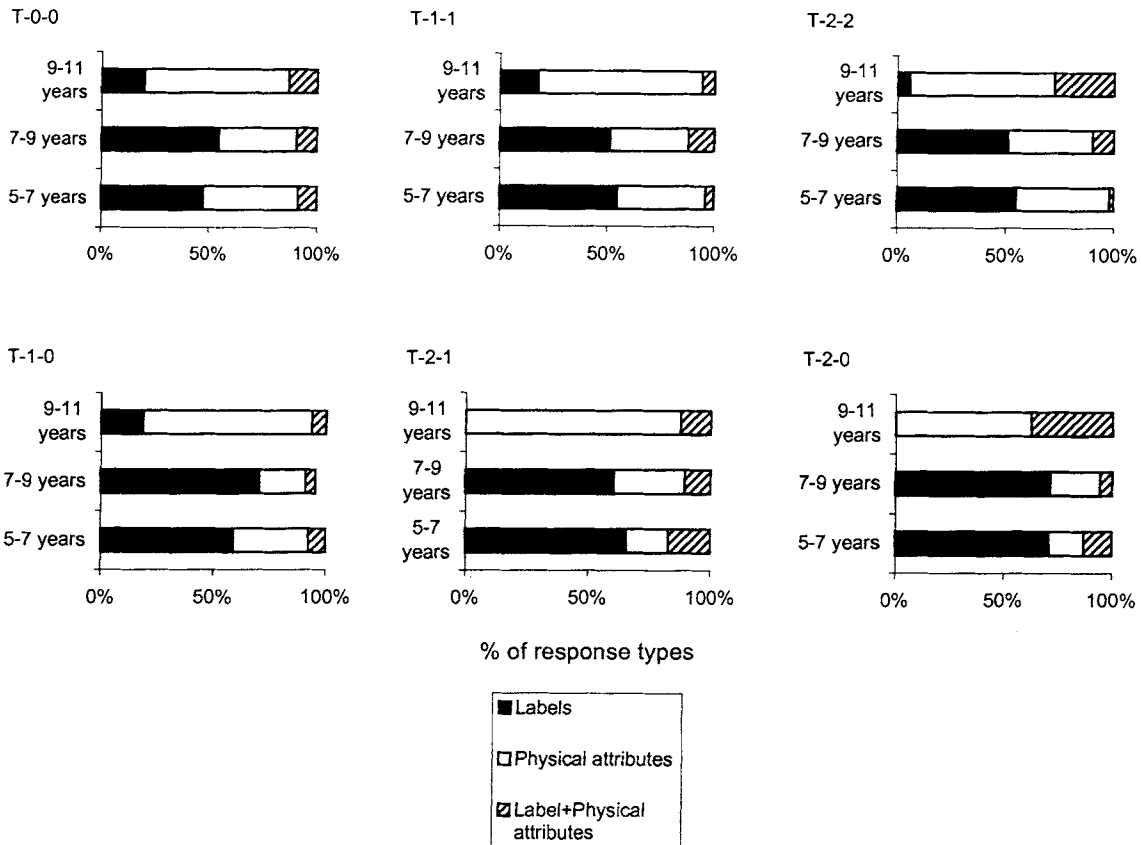
Figure 5. Percentages of different types of explanations in the label condition by age and stimulus pattern. T = target.

spaceships." Each picture had a colored dot near the bottom of the picture. The same female experimenter presented children with a triad, pointed to the dots, and asked children to name the colors. Then she asked children to select one test stimulus that was most similar to the target. After the answers were provided, children were asked to explain their choices.

Results and Discussion

In the analysis of the reported data, we used the following strategy. First, we compared frequencies of label-based choices (derived from Experiment 1) versus dot-based choices (derived from Experiment 2). However, this analysis, although allowing the comparison of the importance of labels and dots, did not allow us to estimate effect sizes. Therefore, we conducted another type of analysis that involved estimating weights of labels and weights of dots.

Frequencies of label-based choices and dot-based choices are presented in Figure 6. Data in Figure 6 indicate that across the 18 trials (3 trials by six stimulus-pattern conditions) the proportion of label-based choices was higher than that of dot-based choices. Recall that each of the six stimulus patterns was presented three times for a total of 18 trials, Mann-Witney $U(1, 18) = 64$, $p < .003$. Note that in each stimulus-pattern condition, children were more likely to make label-based choices than they were to make dot-based choices. This tendency suggests that children's reliance on labels does not stem solely from their lack of confidence or their inability to make a firm judgment when visual attributes did not allow them to make such judgments.

Although data in Figure 6 indicate that labeled patterns were selected over nonlabeled patterns significantly more often than dotted patterns were selected over nondotted patterns, we deemed it necessary to evaluate the magnitude of the effects of labels and
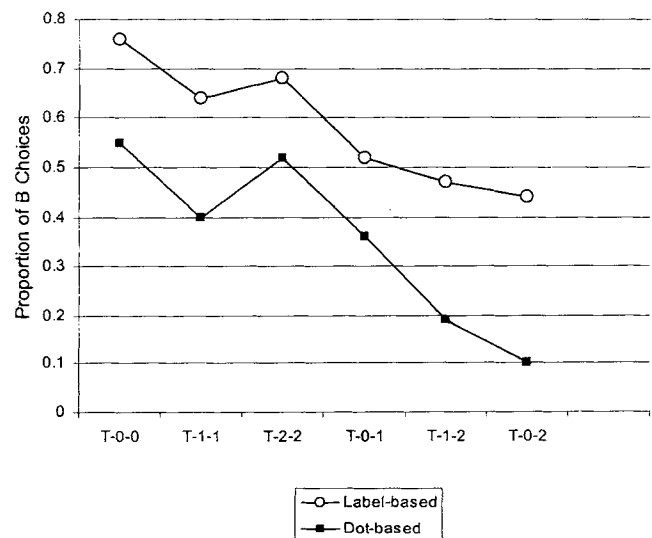

Figure 6. Proportions of label-based versus dot-based choices. T = target.

dots on young children's similarity judgment. To do so, we estimated weights of labels and dots (S_{label} vs. S_{dot}) in the following manner. First, we used the control data of Experiment 1 to estimate the weights of a mismatch (S) of visual attributes. The average weight of visual attributes was equal to 0.48, which was quite close to our a priori assumption of $S_{vis.attr} = 0.5$. Then, for each trial within each stimulus-pattern condition, we substituted $P(B)$ in Equation 7 with observed proportions of selecting Test Stimulus B over Test Stimulus A, thus calculating the value of $Sim(T, A)$. After that, we fitted values of $Sim(T, A)$ and $S_{vis.attr}$ into Equation 6, thus estimating weights of labels or dots. Recall that Equation 6 linked together these variables and parameters:

$$Sim(T, A) = S_{label/dot}S_{vis.attr}^{N-k}$$

These weights aggregated across the 18 trials are presented in Figure 7. The estimates suggest that although labels contribute strongly to similarity judgment ($S < 0.5$), contribution of dots is negligible ($S > 1$). Recall that if the value of S of a feature is close or more than 1, this feature makes no contribution to similarity judgment.

In short, the results of Experiment 2 indicated that the contribution of labels in children's similarity judgment is significantly greater than the contribution of dots that were task-irrelevant stimuli. Therefore, these data allow us to rule out a possibility that young children rely on labels because of their inability or unwillingness to ignore task-irrelevant information or their inability to make firm, confident judgments.

Experiment 3

Recall that the goal of this experiment was to test the cross-modality hypothesis. As noted earlier, it is possible that the reported effects of labels stem not from their representing a different modality but from their representing the language. If this argument is true, then analogs of labels that have linguistic properties (i.e., they denote objects) but that are presented without sound (e.g., signs of a sign language) should have effects similar to those of labels. Should effects of signs be different from those of labels, however, there would be strong support for the cross-modality hypothesis.
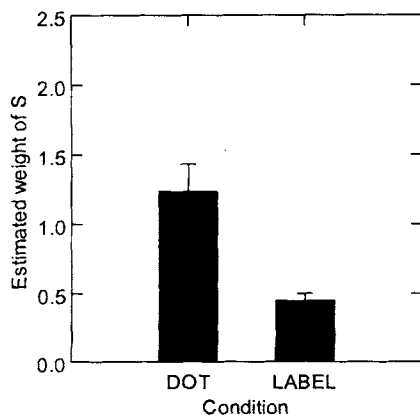

Figure 7. Estimated value of S_{labels} and S_{dots}. S_{label} = weight of a mismatch of labels; S_{dot} = weight of a mismatch of dot's color.

Method

Participants. A group of 26 children participated in the study. These children represented two age groups. The first group consisted of 14 six-to-seven-year-olds ($M = 6.5$ years, $SD = 0.7$; 7 boys and 7 girls). The second group consisted of 12 nine-to-eleven-year-olds ($M = 9.8$ years, $SD = 0.61$; 7 boys and 5 girls). These participants were recruited from an elementary and a middle school located in middle-class suburbs of Columbus, Ohio. All the participants were Caucasians.

Materials. The experiment had the same materials as those used in Experiment 1 except that in the reported experiment, labels were substituted with signs of a sign language. These signs were referred to as "words of alien languages." As in Experiment 2, the reported experiment did not have a control condition.

Procedure. The procedure was identical to that of Experiments 1 and 2. The only significant way in which this experiment was different from the first two was that in the warm-up condition for this experiment, children were familiarized with signs and were told that these signs indicate words. The researcher asked children if they saw signs of a sign language in the past. Then she read the following script:

> I am going to show you some pictures of aliens and tell you where they are from. These aliens use languages different from yours. Instead of using words, they use signs. For example, when we want to name this thing [point to table], we say "a table," but these aliens use a sign [the sign was presented at this time] to name it. In other words, instead of saying "a table," they just use this sign [the sign was shown again]. They use signs all the time when they talk to each other. We know some of their signs. For example, we know the signs that name planets these aliens come from. For example, this alien [show a picture of an alien that will not be used in the experiment] comes from the planet "Gula." Instead of saying that he is "a Gula," he shows this sign [the sign was shown at this time]. Do you understand how they use these signs?

Note that in the course of experiment, signs were never verbalized again.

As in the first two experiments, one of the experimenters (Ya-Fen Lo) presented children with a triad. After that she introduced two signs that were referred to as "words of the alien's languages that mean planets these aliens come from." After that, she asked a child to repeat the signs. Then she asked the child to select one test stimulus that was most similar to the target. After the answers were provided, children were asked to explain their choices.

Results and Discussion

In presenting results of this experiment, we first compare frequencies of label-based choices (derived from Experiment 1) with sign-based choices (derived from Experiment 3). After that, we compare estimated weights of labels with estimated weights of signs. These estimates were derived in a manner identical to that of Experiment 2.

Frequencies of label-based choices and sign-based choices by age group are presented in Figure 8. Data in the figure indicate that for the younger group, across the stimulus-pattern conditions the proportion of label-based choices was significantly higher than that of sign-based choices, Mann-Whitney $U(1) = 264$, $p < .001$. These differences were not significant for the older group, Mann-Whitney $U(1) = 199.5$, $p = .234$

As in Experiment 2, we also deemed it necessary to evaluate the magnitude of effects of labels and signs on young children's similarity judgment. To do so, we estimated weights of labels and signs (S_{label} vs. S_{sign}). These weights aggregated across the 18
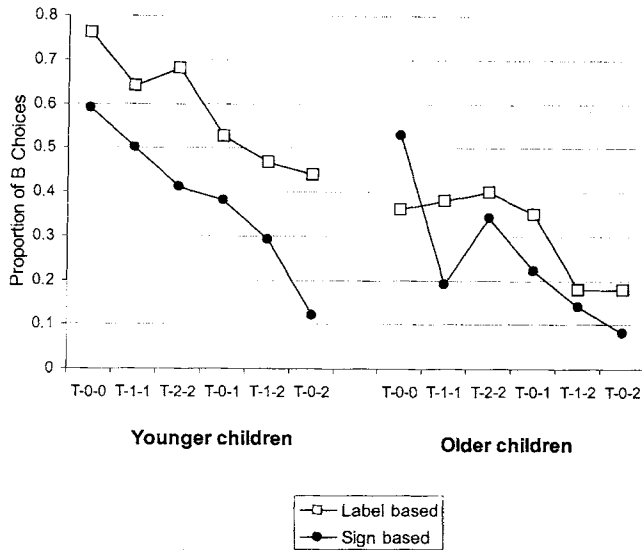
Figure 8. Proportions of label-based versus sign-based choices by age group. T = target.

trials are presented in Figure 9. The estimates suggest that for younger children, labels contributed strongly to similarity judgment ($S < 0.5$), whereas the contribution of signs was negligible ($S > 1$). For older children, the contribution of both labels and dots was negligible ($S > 1$).

Results of this experiment indicated that the contribution of labels in young children's similarity judgment is significantly greater than the contribution of signs. For older children, neither labels nor signs contributed significantly to their similarity judgment. These data provide support for the cross-modality hypothesis. Recall that according to the hypothesis, the effects of labels stem not from their representing the language but from their representing a different modality. Indeed, although labels represent the language plus a different modality, signs represent the language plus the same modality.

Are there any alternative explanations of these findings? Of course, it could be argued that younger children merely do not understand that signs represent the language (Namy & Waxman, 1998) and, as a result, treat them as task-irrelevant information. However, this argument seems less plausible than the cross-modality hypothesis if the latter is taken in conjunction with prior findings that nonlinguistic auditory stimuli generated effects similar to those of labels (Roberts, 1995; Roberts & Jacob, 1991). In addition, during the warm-up phase, all children reported that they saw a sign language in the past, and they had no difficulties answering warm-up questions and translating signs into English. However, it is possible that in addition to the cross-modality of spoken language, there are some phonological and prosodic aspects of human spoken language that attract children's attention (cf. Balaban & Waxman, 1997). Balaban and Waxman demonstrated that auditory stimuli (both spoken words and nonlinguistic tones) have effects on infants' categorization; however, the effects of spoken language were slightly greater than the effects of nonlinguistic tones.

General Discussion

The results of the three reported experiments may be summarized as follows. First, labels strongly contribute to the similarity judgment of young children. Second, labels contribute in a quantitative manner, and this contribution varies with the number of overlapping attributes and with age. Third, the contribution of labels is unlikely to stem from children's inability or unwillingness to ignore task-irrelevant information, because children can ignore other task-irrelevant features, such as colored dots and signs. Finally, the contribution of the labels was much larger than that of nonauditory linguistic entities, such as signs of a sign language.

These findings fit the hypotheses well and support all critical predictions presented in Table 2. As predicted, differences between the label and no-label conditions were pronounced in younger children, whereas they were absent in the older children. Furthermore, the importance of the label decreased with age, and the label contributed to the similarity judgment not in an all-or-nothing manner but in a quantitative manner. In addition to this qualitative fit, the data represented a reasonable quantitative fit for theoretical predictions. In particular, predictions derived from the theoretical model of labels as attributes presented in Equations 6 and 7 correlated highly ($r = .87$) with the observed data. Experiments 2 and 3, in conjunction with prior findings on the importance of nonlinguistic auditory stimuli (Roberts 1995; Roberts & Jacob, 1991), provide additional support for the cross-modality hypotheses. Whereas results of Experiment 2 suggest that effects of labels are unlikely to stem from children's inability to ignore task-irrelevant information, results of Experiment 3 indicate that effects of labels may not stem from the fact that they represent language. Finally, results of Experiments 2 and 3 indicate that whereas for younger children, weights of labels were higher than those of other attributes, weights of dots and signs were lower, thus supporting the assumption that different attributes may have different weights in similarity judgment.

These results fit both assumptions underlying the model: the assumption of label as attribute and the assumption that different attributes may have different weights. The results support a possibility of a low-level mechanism that could steer similarity judgment, grouping, and acquisition of categories in childhood. Indeed,
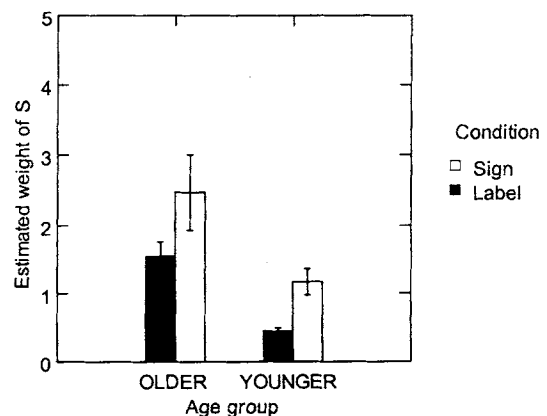

Figure 9. Estimated value of S_{label} and S_{sign} by age group. S_{label} = weight of a mismatch of label; S_{sign} = weight of a mismatch of signs.

if labels are considered to be attributes that have larger weights than other attributes, then features that correlate with labels are more likely to be attended to. These simple properties of labels seem to be able to account for several well-established phenomena. For example, if two novel entities share a label and other features, young children are likely to consider them similar and members of the same category (E. Markman, 1989; Waxman & Markow, 1995). If two nonidentical entities have different labels, young children are likely to grant these entities different feature structures, even if they share some features (Gelman & Markman, 1987). It has also been established that younger children (for whom the base rate of familiar categories is low and the base rate of novel categories is high) are likely to identify a novel object as a member of a novel category when the object has a novel label (E. Markman, 1987). Similarly, if a label accompanies an object, the object is more likely to be interpreted as a member of a category than would an object presented without a label (Waxman & Markow, 1995). In addition, E. Markman (1989) demonstrated that when young children are presented with a familiar object, a novel object, and a novel label, they typically interpret the novel label as denoting the novel object. If children do consider the label as an on/off attribute, which can have only one value at a time (e.g., an object typically has one shape and one size, although sometimes it might have more than one color), then they would be reluctant to grant different labels to the same object.

Recall that some of these findings have been captured in Equations 1–4, presented earlier. However, relations denoted by these equations seem to get degraded in the course of development. As the base rate of novel categories decreases and as children acquire homonyms and synonyms, are exposed to predicates that poorly correlate with other features (e.g., a fan of the Chicago Bulls), and acquire subordinate, superordinate, and collateral categories, they accumulate examples that decrease predictive validity of labels. As a result of these developments, attentional weights of labels decrease, thus allowing children to maintain an acquired categorical structure that is relatively stable.

Equations 1–4 allow us to generate several predictions that could lead to further tests of the proposed model. According to Equation 1, young children should be likely to use labels not only in similarity judgment but also in several similar tasks, including inductive predictions about novel objects. For example, in a reversed version of Smith et al.'s (1996) experiment, children could be presented with three unfamiliar objects (artifacts or natural kinds), two of which have overlapping features, and with two other objects that have the same novel label (e.g., *Amlas*). If told that one of the Amlases "can hold pens" (or "has triangular brain cells") and asked which of the two is likely to hold pens (or to have triangular brain cells), young children should base their induction on the shared label. This, however, should not be the case with older children, who should be more likely to look for more informative features. When children cannot rely on perceptual information (e.g., if all the compared objects are sufficiently similar, or if no perceptual information is offered), then both younger and older children should base their induction on the shared label. However, if labels of the compared novel objects do not vary (all the objects have the same label or labels are not provided), then when making inductive predictions, both younger and older children should exhibit strong reliance on overlapping features (see Equation 2).

Equation 3 allows us to make another interesting prediction. Recall that Equation 3 is based on a simple notion that the base rate of novel categories varies with age. Roughly speaking, this base rate is quite high during the acquisition mode and is relatively low during the maintenance mode. If presented with a novel object (an artifact or a natural kind) and a novel label (e.g., *Gubla*) and asked if there are other Gublas in the world (the question that checks if a person wants to grant a category status to an object), younger children should answer "yes," whereas older children should perform at the chance level, or balk at the question. The same regularities should hold for induction questions (e.g., "Suppose that there are other Gublas. This Gubla has thick blood. Do you think that other Gublas also have thick blood?"). If this prediction is correct, then we could answer one question that was implicitly mentioned earlier but that did not get a satisfactory resolution: At what age is there a breaking point such that the base rate of familiar categories is larger than the base rate of novel categories? The probable answer is when the frequency of "yes" responses to the Gubla question drops to chance or below. Note that effects of low base rates of novel categories could be easily eliminated if the Gubla is introduced as an inhabitant of an imaginary or fictitious world, for which the acquired base rates do not hold. In this case, both older and younger children may answer "yes" to the Gubla question.

There are several issues that require further empirical clarification. First, the stimuli (schematic faces with a small number of discrete attributes and artificial labels) represented both the advantages and disadvantages of studying artificial entities. On the one hand, these stimuli were simple, tractable, quantifiable, and not confounded by prior knowledge. On the other hand, it remains unclear whether these stimuli were fully representative of real-life entities (e.g., animals, plants, people, or vehicles). There is evidence that cognitive processes (such as learning, memory, or categorization) involving artificial entities are representative of those involving more natural ones (Estes, 1994); however, these findings are limited to adults.

Second, there are aspects of models of similarity that require further analyses. For example, the extent to which people, when computing similarity, use principles captured by the product-rule model remains unclear. When stimuli include relatively few features (as in our experiments) and these are relatively similar on the larger scale (e.g., they are humanlike faces, or animals, or block-type objects), then the product-rule model yields reasonable predictions. However, it seems unlikely that people perform feature-by-feature computations when compared objects are profoundly dissimilar (e.g., automobile vs. jellyfish). It seems more plausible that in the latter case, people use several of the most salient features to arrive at a categorical distinction between the compared objects and that they stop further comparisons once they identify objects as categorically different. In fact, it has been demonstrated that people use such categorical distinctions (e.g., between letters and numbers) when performing visual search (Schneider & Shiffrin, 1977). In addition, there is evidence that identity on one dimension may weigh more than a close proximity on a variety of dimensions (e.g., Smith, 1989). For example, younger children may tend to group a black circle with a gray ellipse rather than with a white circle or a black rectangle. Older children, however, may tend to group in accordance with dimensional identity (e.g., putting a black circle with a white circle rather than with a gray

ellipse). However, if there is categorical identity, then even for older participants, overall similarity may override dimensional identity (e.g., blue sedan and yellow tractor vs. blue jellyfish). However, the existing models of similarity do not specify conditions under which people attend to overall similarity as opposed to those conditions under which they prefer dimensional identity.

There is another aspect of similarity that is conspicuously absent from most formal models—the distinction between elementary and relational features (but see Goldstone et al., 1991, and Medin, Goldstone, & Gentner, 1990, for notable exceptions). Which objects would be judged to be more similar: those that share all and only elementary features (e.g., shape, size, and color) or those that share relational features (e.g., spatial symmetry of objects or correlations among elementary features)? Previous studies have demonstrated that under certain conditions, infants, young children, and adults could be more sensitive to relational than to elementary features (Kotovsky & Gentner, 1996; Medin et al., 1990; Younger & Cohen, 1983; Younger & Fearing, 1998).

Third, the cross-modality hypothesis requires additional examination. Although the importance of auditory input in detecting similarities between visual patterns has been demonstrated (Roberts, 1995; Roberts & Jacob, 1991), it has been also argued that human spoken language may contribute to this process above and beyond nonlinguistic auditory stimuli (Balaban & Waxman, 1997). Therefore, it seems necessary to compare effect sizes of spoken language and those of auditory stimuli with effect sizes of cross-modal stimuli and those of same-modal stimuli. For example, Balaban and Waxman reported large effect sizes (approximately 0.66 averaged across experiments) of the cross-modality, whereas we could not locate reports on effect sizes of spoken language. Small effect sizes of spoken language as compared with auditory stimuli could bring additional support for the cross-modality hypothesis.

Finally, the proposed model also needs a more detailed developmental analysis. Because there is no precise theory specifying age differences in similarity judgment, the studied groups were formed in a rather arbitrary manner. Although adolescents could be close to a developmental "upper bound," kindergartners and first graders are definitely not close to the "lower bound." For example, in the majority of previous studies, participants were significantly younger than kindergartners. However, this limitation could be easily overcome in future research. If our general theoretical prediction is correct, the weight of the label should be even larger in infants, toddlers, and preschoolers than it is in 6–7-year-olds, unless there is a dissociation between labels and feature sets (see Davidson & Gelman, 1990, for effects of a double dissociation).

In short, the proposed model of label as attribute provides a good qualitative account of the data, and it lends itself to precise quantitative predictions. Furthermore, it allows us to make a set of testable predictions that will be examined in future studies. In addition, it can also account for some well-established phenomena, including children's tendency to focus on labels when performing categorization, induction, and similarity judgment tasks (Balaban & Waxman, 1997; Gelman & Markman, 1987; E. Markman, 1989; Waxman & Markow, 1995). However, additional research is needed to examine the validity of the model with respect to naturally occurring phenomena, to carry out a more precise developmental analysis, and to increase the quantitative precision of theoretical predictions. Such research should include larger data

sets, more diverse stimuli, wider age ranges, and more dense developmental observations.

Overall, the reported experiments allow us to conclude that under these conditions, children consider labels to be an attribute of an object; labels contribute to similarity judgment in a quantitative manner; in addition, the relative weights of labels decrease with age—they are greater for younger children and smaller for older ones.

References

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Backscheider, A. G., & Gelman, S. A. (1995). Children's understanding of homonyms. *Journal of Child Language, 22,* 107–127.

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology, 64,* 3–26.

Brown, A. (1990). Domain-specific principles affecting learning and transfer in children. *Cognitive Science, 14,* 107–133.

Estes, W. K. (1994). *Classification and cognition.* New York: Oxford University Press.

Davidson, N. S., & Gelman, S. A. (1990). Inductions from novel categories: The role of language and conceptual structure. *Cognitive Development, 5,* 151–176.

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

Gelman, S. A., & O'Reilly, A. W. (1988). Children's inductive inferences within superordinate categories: The role of language and category structure. *Child Development, 59,* 876–887.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essence: Early understandings of the non-obvious. *Cognition, 38,* 213–244.

Gentner, D. (1978). A study of early word meaning: What looks like a jiggy but acts like a zimbo? *Papers and Reports on Child Language Development, 15,* 1–6.

Gentner, D., & Markman, A. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52,* 45–56.

Gentner, D., & Rattermann, M. (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). New York: Cambridge University Press.

Gick, M., & Holyoak, K. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–365.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology, 23,* 222–262.

Goodman, N. (1992). Seven strictures on similarity. In M. Douglas & D. Hull (Eds.), *How classification works* (pp. 13–23). Edinburgh, Scotland: Edinburgh University Press. (Original work published 1972)

Gopnik, A., & Meltzoff, A. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Goswami, U. (1992). *Analogical reasoning in children.* Hove, England: Erlbaum.

Goswami, U., & Brown, A. (1989). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition, 35,* 69–95.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67,* 2797–2822.

Landau, B., Smith, L., & Jones, S. (1997). Object shape, object function, and object name. *Journal of Memory & Language, 36,* 1–27.

Lopez, A., Gelman, S., Gutheil, G., & Smith, E. (1992). The development of category-based induction. *Child Development, 63,* 1070–1090.

Lowenstein, J., & Gentner, D. (1998, August). *Relational language facilitates analogy in children.* Paper presented at the 20th Meeting of the Cognitive Science Society, Madison, WI.

Mandler, J. M. (1997). Development of categorisation: Perceptual and conceptual categories. In G. Bremner, A. Slater, & G. Butterworth (Eds.), *Infant development: Recent advances* (pp. 163–189). Hove, England: Psychology Press.

Mandler, J. M., Bauer, P. J., & McDonough L. (1991). Separating the sheep from the goats: Differentiating global categories. *Cognitive Psychology, 23,* 263–298.

Mandler, J. M., & McDonough, L. (1996). Drinking and driving don't mix: Inductive generalization in infancy. *Cognition, 59,* 307–335.

Markman, A. B., & Gentner, D. (1997). The effects of alignability on memory. *Psychological Science, 8,* 363–367.

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

Markman, E. M. (1987). How children constrain the possible meanings of words. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 255–287). Cambridge, MA: Cambridge University 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. (1975). A theory of context in discrimination learning. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9, pp. 263–314). New York: Academic Press.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science, 1,* 64–69.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254–278.

Namy, L. L., & Waxman, S. R. (1998). Words and gestures: Infants' interpretations of different forms of symbolic reference. *Child Development, 69,* 295–308.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97,* 185–200.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

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

Roberts, K. (1995). Categorical responding in 15-month-olds: Influence of the noun-category bias and the covariation between visual fixation and auditory input. *Cognitive Development, 10,* 21–41.

Roberts, K., & Jacob, M. (1991). Linguistic vs. attentional influences on nonlinguistic categorization in 15-month-old infants. *Cognitive Development, 6,* 355–375.

Roediger, H. L. (1990). Implicit memory: Retention without remembering: *American Psychologist, 45,* 1043–1056.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Ross, B. H. (1989). Remindings in learning and instruction. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 438–469). New York: Cambridge University Press.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1–67.

Smith, L. (1989). A model of perceptual classification in children and adults. *Psychological Review, 96,* 125–144.

Smith, L., & Heise, D. (1992). Perceptual similarity and conceptual structure. In B. Burns (Ed.), *Advances in psychology: Vol. 93. Percepts, concepts and categories: The representation and processing of information* (pp. 233–272). Amsterdam: North-Holland.

Smith, L. B., Jones, S. S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism. *Cognition, 60,* 143–171.

Soja, N. N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. *Cognition, 38,* 179–211.

Sperber, R. D., Davies, D., Merrill, E. C., & McCauley, C. (1982). Cross-category differences in the processing of subordinate/superordinate relationships. *Child Development, 53,* 1249–1253.

Thorndike, E. L. (1913). *Educational psychology: Vol. 2. The psychology of learning.* New York: Teachers College Press.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Vygotsky, L. S. (1962). *Thought and language.* Cambridge, MA: MIT 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,* 557–302.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43,* 337–375.

Xu, F. (1998, August). *Weakly Whorfian? The role of labeling in acquiring kind concepts in infancy.* Paper presented at the 20th Meeting of the Cognitive Science Society, Madison, WI.

Younger, B. A., & Cohen, L. B. (1983). Infant perception of correlations among attributes. *Child Development, 54,* 858–869.

Younger, B., & Cohen, L. (1986). Developmental change in infants' perception of correlations among attributes. *Child Development, 57,* 803–815.

Younger, B. A., & Fearing, D. D. (1998). Detecting correlations among form attributes: An object-examining test with infants. *Infant Behavior & Development, 21,* 289–297.