Psychological Bulletin
2001, Vol. 127, No. 6, 827–852

0033-2909/01/$5.00 DOI: 10.1037//0033-2909.127.6.827

Necessity and Natural Categories

Lance J. Rips
Northwestern University

Our knowledge of natural categories includes beliefs not only about what is true of them but also about what would be true if the categories had properties other than (or in addition to) their actual ones. Evidence about these beliefs comes from three lines of research: experiments on category-based induction, on hypothetical transformations of category members, and on definitions of kind terms. The 1st part of this article examines results and theories arising from each of these research streams. The 2nd part considers possible unified theories for this domain, including theories based on ideals and norms. It also contrasts 2 broad frameworks for modal category information: one focusing on beliefs about intrinsic or essential properties, the other focusing on interacting causal relations.

It's common ground in linguistics, artificial intelligence (AI), and philosophy that our knowledge of natural categories includes information that is resistant to exceptions. Linguists, for example, have described generic sentences, such as *Lions have manes,* as ones that are true, despite the existence of obvious and sometimes numerous exceptions (such as female lions and immature male lions; see Krifka et al., 1995). Likewise, research on nonmonotonic logic in AI has sought systems that can reason with such sentences without making mistakes or becoming inconsistent when exceptions arise (e.g., Ginsberg, 1987). Some theories in the philosophy of language invest everyday concepts such as *lion* with a status that allows them to play a role in counterfactual conditionals, such as *If Calvin were a lion, he'd have a mane* (e.g., Brandom, 1988, 1994). Not only does our knowledge of categories withstand exceptional current circumstances, it stands as well in merely possible circumstances that we have not experienced.

Cognitive psychology, however, has mostly treated beliefs about categories in terms of what's normal or usual rather than in terms of what's lawlike or exception resistant. Early theories of perceptual categorizing (e.g., Posner & Keele, 1968; Reed, 1972) emphasized the role of prototypes, consisting of average values of category members along their physical dimensions. According to these theories, if people have to classify, for example, schematic faces into two previously identified sets, they mentally compute a prototype for each set, where the prototype specifies the average values of the members of that set on dimensions such as width of mouth, length of nose, and distance between eyes. To decide which set a novel face belongs to, people then determine the distance between the new face and each of the category prototypes. Finally, people assign the new face to the set whose prototype is closest to this new item.

National Science Foundation Grants SBR-9514491 and SES-9907414 supported this research. Thanks go to Serge Blok, Douglas Medin, Daniel Osherson, and Steven Sloman for comments on an earlier version of this article and to Beth Proffitt for conversations about the material in the first part of the article.

Correspondence concerning this article should be addressed to Lance J. Rips, Psychology Department, Northwestern University, 2029 Sheridan Road, Evanston, Illinois 60208. Electronic mail may be sent to rips@northwestern.edu.

The importance of normal or average values of category members persists in many cognitive theories of everyday categories, such as lions or pajamas. In Eleanor Rosch's well-known theory (e.g., Rosch, 1978; Rosch & Mervis, 1975), membership in these categories depends on the typicality of an instance with respect to the category. Typicality of the instance depends, in turn, on how many of its stimulus values the instance shares with members of the target category and how few values it shares with members of rival categories. For example, the best examples of lions—the most typical ones—are those that have properties that are most widespread among lions (and the least widespread among cougars, cheetahs, tigers, etc.). On this view, then, both typicality and category membership come down to possessing properties (values of attributes, such as having a tawny color) that are common in a census of the target category (see A. Tversky, 1977, for a similar view of typicality). Although Rosch (1978) held that there may be no single prototype for everyday categories, she nevertheless believed that these categories depend on the prevalence of properties among their instances.

Current theories have often taken over this view of everyday categories as based on average values, although the mental representations that contain them tend to be more complex. For example, Hampton (1995a) retained the notion of a prototype as a "generalization or abstraction of some central tendency, average or typical value of a class of instances falling in the same category" (p. 104), and Smith, Osherson, Rips, and Keane's (1988) prototypes are composed of attribute–value combinations, with each value weighted (in part) by the subjective frequency of the value among category instances. To be sure, there are psychological theories that do not rely on prototypes, but many of these alternatives also appeal to the properties of (samples of) existing category members. For instance, in exemplar theories of categories (e.g., Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986), decisions about membership depend on the similarity of new items to previously encountered and remembered instances. Exemplar models do not compile average values or distributions of properties for a category in the way prototype theories do, but to an even greater extent, they are tethered to a sample of actual category members.

I examine other psychological theories later in this article, but for now the issue is this: If people's concepts of natural categories are based on their surveys of members of these categories, as they seem to be in prototype or exemplar theories, then what underwrites their confidence that these concepts resist exceptions, generalize to novel instances, and support counterfactual conditionals? To see the difficulty here, compare lions to a nonnatural category, such as things in offices weighing between 40 and 50 pounds. An inventory might convince you that such things are typically beige in color and rectangular in shape. But you would probably be more hesitant to attribute beige and rectangular to arbitrary new members of this category than to attribute tawny color and lionlike shape to new members of the lion category. Similarly, you would never suppose that if Calvin were to weigh 45 pounds and were to step into an office, he would be beige; but you might well think that if Calvin were a lion he'd be tawny and lion shaped. It is entirely possible that on-the-spot perceptual recognition of objects as members of natural kinds often depends on average or typical stimulus values. As the questions just raised make clear, however, once people begin to consider the role knowledge of natural kinds plays in other forms of thought, they are forced to take into account these kinds' *modal* properties—properties that the members of the kind *might* have or *must* have across (possibly counterfactual) circumstances. Since my interest is in modal properties here, I focus on the functions of natural kinds in reasoning rather than in perception.

In the first part of this article, I look at some psychological evidence that pins down the modal qualities of natural kinds (e.g., daisies, lions, copper); in the second part, I examine some approaches that may have a chance of explaining these qualities. Although natural kinds have a central place in psychological and philosophical theories of concepts and although there is a great deal of research about them, their modal properties are less obvious and less well understood. The aim of this article is to grapple with the question of how seemingly objective categories such as lions could have properties that extend beyond the set of their actual members.

Modal Characteristics of Natural Categories: Psychological Evidence

In recent cognitive studies of natural categories, there seem to be three main ways in which beliefs about these categories could be said to be modal: First, natural categories appear to govern people's ideas about the distribution of their relevant properties, even in the face of contradictory perceptual evidence. Second, these categories dictate conditions under which individuals belong to the category, again despite perceptual evidence. And third, people think there are determinate ways of resolving questions about category membership, although they may have no personal knowledge of what these tests might be. I review these experiments here by way of finding out the nature of these beliefs' modal character. Although I briefly examine children's ideas about these matters, the main goal is to determine the final state of these beliefs among adults.

Natural Categories and Induction

People think of a natural category as governing the properties of its members. For example, they think that certain biological properties of lions, such as having lungs, having four legs, or having a specific genetic makeup, will tend to be potentially true of all members of the lion category. Thus, if they learn that a new biological property is true of a particular lion, they are likely to think that other lions have the property as well. If you learn that Leigh has Type K blood serum, for example, you are likely to think that other lions also have it. Type K blood is presumably a type of blood, even though it's not a type you've heard of. Since types of blood are the sorts of properties that tend to run along biological lines, you are willing to generalize them to other lions. You could be wrong. Maybe lions, like people, have more than one blood type. But your willingness to generalize is an important aspect of your knowledge of categories. Of course, not all properties generalize across natural kinds: Leigh's particular pattern of cuts and bruises is not the sort of property that is likely to be true of other lions. The power of natural kinds to guide generalization therefore depends on the type of relation that holds between property and kind.

Category-Based Inductive Inferences in Children

Even toddlers go along with generalizing by kind. As early as 14 months, infants generalize an animal activity more often from one toy animal to another than from an animal to an artifact (Mandler & McDonough, 1998). For example, after they have seen an experimenter demonstrate a dog drinking, they tend to imitate the drinking more often with a lamb than with a train engine. When the activity is not specific to animals, however, they generalize about equally to another animal as to an artifact (e.g., they generalize getting cleaned with a sponge from the dog to the train engine about as often as to the lamb).

Mandler and McDonough found little sensitivity to distinctions within the category of land animals. But by 2 or 3 years, children more often generalize familiar properties to novel instances of the same lower level category than to novel instances of other categories at the same level (Gelman & Coley, 1990; Waxman, Lynch, Casey, & Baer, 1997). For example, Waxman et al. told children that a pictured animal had a specific (but unpictured) property and then asked whether the same property was true of other animals. The children learned, for instance, that a particular collie had the property of "helping us take care of sheep." The children then had to decide whether other collies, other dogs (e.g., setters), and other animals (e.g., caribou) also had this property. The results indicated that children generalized the property to other dogs more often than to nondog animals. Training on contrasting properties of subcategories (e.g., setters "help us find birds," whereas samoyeds "help us pull sleds") further restricted the range of the children's generalization. Somewhat older children (4-year-olds) are usually willing to generalize unfamiliar biological properties (e.g., *having cold blood*) by category, even when perceptual appearance is placed in direct opposition to category membership (Gelman & Markman, 1986). These children, for example, preferred generalizing the property *having cold blood* from one dinosaur (a pictured brontosaurus) to a second dinosaur (triceratops) over generalizing *having warm blood* from a rhino to the same triceratops, even though the picture of the

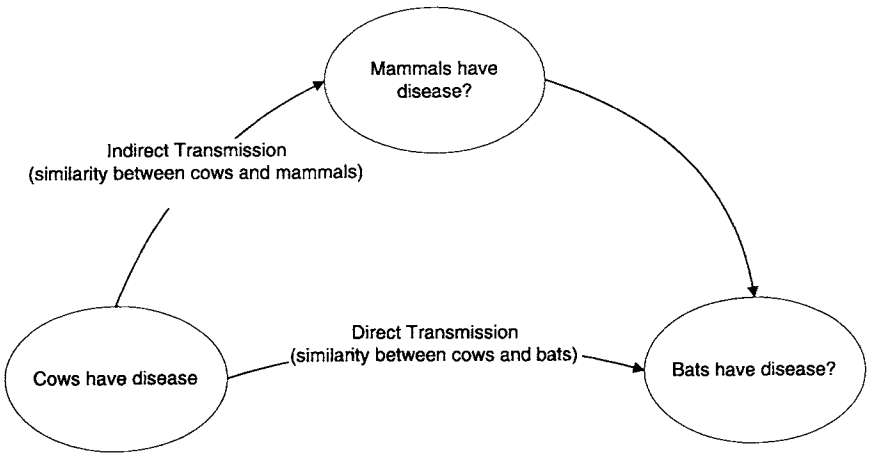

Figure 1. An example of the similarity-coverage model (Osherson et al., 1990) as applied to the problem

Cows have disease X.

Bats have disease X.

The model computes the strength of the conclusion as a weighted average of the direct similarity between cows and bats (direct transmission route) and the similarity between cows and all mammal species (indirect transmission route).

triceratops looked more similar to the rhino than to the brontosaurus.[1]

This, of course, does not mean that children always generalize properties in an adultlike way. Carey (1985, chap. 4) found that 4-year-olds are much more apt to generalize unfamiliar properties (e.g., *having a spleen inside*) from people to other familiar animals, such as dogs, than from dogs to people. Carey took this result to indicate that younger children's knowledge of the animal domain is initially organized in terms of their beliefs about specifically human activities rather than in terms of biological characteristics. Carey (1985) and Gelman and O'Reilly (1988) also found that 4-year-olds are less willing than early grade-school children to generalize unfamiliar properties from one member of a natural kind to a member of a second kind within the same superordinate category. For example, the younger children were more hesitant to generalize the property *has leukocytes all through it* from a dog to a horse than were the older children. Evidence is somewhat inconsistent on whether preschool children recognize that natural kinds are more likely than artifacts to promote generalization (Gelman, 1988; Gelman & O'Reilly, 1988), presumably because these children are just mastering the relevant knowledge. In sum, these results suggest that 4-year-olds understand animal and plant species as supporting inductive generalization to some extent, and they have a rough idea of which properties generalize in this way and which do not. Nevertheless, children 4 or younger apparently have no clear sense of the mechanisms that support generalization over species or of the differences between these mechanisms and those at work in artifacts. Finally, even 8-year-olds have difficulty recognizing the importance of sample size and sample variability in induction based on natural categories (Gutheil & Gelman, 1997).

Category-Based Inductive Inferences in Adults

Early theories of category-based inductive inference in adults (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rips, 1975) consisted of a two-part process: a direct and an indirect transmission of the target property. According to these theories, when people are told, for example, that cows have a novel disease and are asked to estimate the likelihood that bats also have the disease, they consider both the similarity between cows and bats (direct transmission) and the similarity or typicality of cows with respect to mammals in general (indirect transmission). These two routes to an inductive conclusion appear in Figure 1. The first route accounts for the finding that people judge it more likely that the disease will generalize from cows to horses than from cows to bats. The second route accounts for the fact that people judge it more likely that the disease will generalize from cows (a typical mammal) to bats (an atypical mammal) than from bats to cows.

The similarity-coverage model. Osherson et al. (1990) elaborated this theory in their similarity-coverage model to describe simultaneous generalization from several different categories. This model is able to predict, for example, how likely people think the conclusion of Argument 1 is, given the truth of its premises:

(1) Cows have Vitamin Z.
Lions have Vitamin Z.
Mice have Vitamin Z.

Bats have Vitamin Z.

Applied to an argument of this sort, the similarity-coverage model computes direct transmission as the similarity of the conclusion item (bats) to the most similar of the premise categories. In Argument 1, bats are more similar to mice than to either cows or lions, so the similarity between mice and bats determines the degree of generalization due to the direct route. The model com-

[1] Gelman and Markman (1986) pretested properties like *having cold blood* to make sure their participants didn't already know, for example, that the triceratops had cold blood, so the properties were unfamiliar to these participants.

putes indirect transmission in a related way, by first determining the smallest superordinate category containing all the categories in the premises and conclusion—presumably, mammal in the case of Argument 1. The model then finds the average similarity between the premise categories and each species of mammal known to the participant. Thus, indirect transmission of Vitamin Z would be due to the average similarity of cows, lions, and mice, on one hand, to horses, pigs, bears, and all other mammal species on the other. (The model again calculates the joint similarity of cows, lions, and mice to another species as the similarity of that species to the most similar of those three premise categories.) The overall strength of the argument will be a weighted average of the similarity from direct and indirect routes.

Osherson et al. (1990) show that the similarity-coverage model can explain many phenomena associated with arguments such as 1 that contain biological-seeming, but unfamiliar, properties.[2] In addition to the similarity and typicality results just discussed, the model's indirect transmission route can also explain why participants judge arguments like 1 to be stronger than matched arguments such as 2:

(2) Cows have Vitamin Z.
Horses have Vitamin Z.
Mice have Vitamin Z.

Bats have Vitamin Z.

According to the model, this phenomenon, which Osherson et al. call *premise diversity,* is due to the fact that cows, lions, and mice, as a group, are more similar to other mammal species than are cows, horses, and mice; the first group better spans the range of mammals than the second. It is more likely, therefore, that mammals in general will have Vitamin Z given the premises of Argument 1 than those of Argument 2. The model also applies to arguments in which the conclusion category is at a higher level than the premise categories—for example, those in which *mammals* substitutes for *bats* in Arguments 1 and 2. For these *general* arguments, the direct and the indirect routes collapse; both become just the average similarity of the premise categories to all mammal species. Thus, the model also predicts that Argument 1 will be stronger than Argument 2 when *mammals* appears in place of *bats* in the conclusions.

Perceived similarity, direct and indirect, is the engine of the similarity-coverage model. Sloman (1993) has shown that many of the phenomena that the similarity-coverage model accounts for can also be explained by a single-route similarity theory. This theory represents each of the categories as a set of predicates or features (e.g., *living, has a mane,* and *roars* in the case of lions), and it predicts the strength of arguments such as 1 and 2 to be the proportion of the conclusion category's features that are included among those of the premise categories. For example, this *feature-based* theory explains premise diversity (e.g., the greater strength of Argument 1 than 2) on the assumption that more diverse premise categories will usually contain more of the conclusion category's features than will less diverse premise categories. This one-route similarity model has some strengths and weaknesses relative to Osherson et al.'s (1990) two-route model, but because similarity is the driving force in both models, I consider them together here.

Questions about the similarity-coverage theory. Although the similarity-coverage model is successful in unifying a large set of findings, it runs into a number of difficulties (as does that in Rips, 1975). The most obvious of these is that the model is completely insensitive to the type of property being projected from the premises to the conclusion. It is for this reason that in earlier work, Osherson, Smith, and Shafir (1986) had declared that the similarity idea "oversimplifies the psychology of argument strength" (p. 220).

Some results of Heit and Rubinstein (1994) illustrate the problem (see Osherson et al., 1986, for an earlier example, and Ross & Murphy, 1999, for a later one). These investigators used triples of animal categories, such as bears, tunas, and whales, in which the first and third categories shared certain anatomical properties (e.g., mammalian ones) and the second and third shared behavioral properties (e.g., swimming). Heit and Rubinstein used these category triples to construct arguments, such as the examples in Arguments 3 and 4, varying both the anatomical or behavioral consistency of the categories and the anatomical or behavioral nature of the predicates:

(3) a. Bears have a liver with two chambers that act as one.

Whales have a liver with two chambers that act as one.

b. Tunas have a liver with two chambers that act as one.

Whales have a liver with two chambers that act as one.

(4) a. Bears usually travel in a back-and-forth or zig-zag trajectory.

Whales usually travel in a back-and-forth or zig-zag trajectory.

b. Tunas usually travel in a back-and-forth or zig-zag trajectory.

Whales usually travel in a back-and-forth or zig-zag trajectory.

A theory based entirely on the similarity between the named categories must predict that participants should select the (a)-items as stronger in both Arguments 3 and 4 or should select the (b)-items as stronger in both. For example, if the combination of direct and indirect routes in Figure 1 yields the result that 3a is a stronger inference than 3b, then participants should also choose 4a over 4b, because these arguments involve exactly the same categories. This is not, however, the result that Heit and Rubinstein obtained. Instead, participants found 3a stronger than 3b but 4b stronger than 4a. The categories in the (a)-items are the ones that share their anatomical natures, whereas the categories in the (b)-items share behavioral natures. The premise category in 3a, bears, is therefore a better predictor of the anatomical property *having a liver with two chambers* than is the premise category in 3b. By contrast, the premise category in 4b, tunas, is a better predictor of the behavioral property *traveling in a back-and-forth trajectory* than is the premise category in 4a.

[2] Osherson et al. (1990) call these properties "blank predicates." People probably interpret them, however, as fictitious or unfamiliar biological properties rather than as purely unknown or nonsense predicates. As Osherson et al. (1990) noted, "blank predicates are recognizably scientific in character" (p. 186). I therefore drop "blank predicate" in favor of "unfamiliar biological predicate" in what follows (as do McDonald, Samuels, & Rispoli, 1996).

The similarity-coverage model attempted to sidestep such problems through the use of fictitious biological properties, such as having Vitamin Z, but related difficulties affect the model even for this type of unfamiliar material. These difficulties become apparent when we try to apply the similarity-coverage theory to some of the results on children's inferences. As we noticed earlier, Gelman and Markman (1986) found that 4-year-olds prefer generalizing an unfamiliar biological property by category to generalizing by perceptual similarity. For example, when the children learned that "this dinosaur" (a brontosaurus) has cold blood and that "this rhinoceros" has warm blood, they ventured that "this [new] dinosaur" (a triceratops) has cold rather than warm blood. Osherson et al. (1990) reconstructed Gelman and Markman's task as involving a choice between two contrary arguments. One argument is as follows: Brontosauruses have cold blood; therefore, triceratops have cold blood. The other argument is this: Rhinos have warm blood; therefore, triceratops have warm blood. The stronger of these two arguments should determine the child's response. Which argument is stronger depends on the relative weight given to the direct and indirect routes in Figure 1. On one hand, the argument from the brontosaurus to the triceratops, according to the model, is mainly warranted by the indirect transmission of cold blood through dinosaur. On the other hand, the argument from the rhino to the triceratops will mainly depend on direct similarity. Thus, the similarity-coverage model is consistent with the result, provided that the children place more weight on indirect than on direct transmission.

But although the similarity-coverage model is consistent with the Gelman and Markman (1986) finding, it doesn't explain the result, as the model provides no reason why indirect transmission should dominate direct transmission in this setting. On an intuitive account, the children are correctly generalizing on the basis of the fact that the property is a biologically relevant one, and brontosauruses and triceratops are in the same lower level biological category (i.e., dinosaurs), whereas rhinos and triceratops are not. This is essentially Gelman and Markman's view (see also Markman, 1989, chap. 5). If the property to be projected is an accidental one, such as *can eat a cupful of food* or *has feet that get cold at night,* then children do not generalize by kind any more often than by perceptual similarity (Gelman & Markman, 1986, Experiment 3). Similarly, children as young as 4 years generalize inborn characteristics but not acquired ones (Springer & Keil, 1989, Experiment 3). But this is just the sort of dependence on properties that pure similarity theories, including the similarity-coverage model, can't handle. Thus, portraying the Gelman–Markman findings as a result of a clash between the direct and indirect routes in Figure 1 restates the conditions of the experiment; it doesn't predict the results.[3]

Finally, in carrying out an inductive inference with an unfamiliar property, people sometimes consider connections between premise and conclusion categories that don't hinge on similarity at all, and this leads to mispredictions for the similarity-coverage approach. For example, Lopez, Atran, Coley, Medin, and Smith (1997) reported that the Itzaj (Mayan natives of Guatemala) do not exhibit the premise diversity effect (the contrast between Arguments 1 and 2 above). For example, Itzaj informants were told that coconut palms and royal palms have one disease and that coconut palms and basket whists have another (where coconuts, royal palms, and basket whists are all palm trees known to the informants). When asked whether all other palms were more likely to have the first disease or the second, the Itzaj informants split their vote, despite the fact that coconuts and basket whists are more diverse than coconuts and royal palms in Itzaj folk taxonomies. As justification for preferring the less diverse argument, Itzaj explained that both coconuts and royal palms are tall trees and are therefore more likely to spread a disease by contact with others in the rain forest. As mentioned earlier, the diversity effect depends on participants using the indirect transmission route in Figure 1, according to the similarity-coverage approach. Since coconuts and basket whists are jointly more similar to other palms than are coconuts and royal palms, indirect transmission predicts more generalization from the former pair. At least some Itzaj, however, prefer to reason with more direct causal connections between category members, short-circuiting the similarity-based method (for related evidence, see also Coley, Medin, Proffitt, Lynch, & Atran, 1999; Medin, Lynch, Coley, & Atran, 1997; Proffitt, Coley, & Medin, 2000).

The gap model. Osherson et al. (1986) were well aware of the limitations of theories based on pure similarity. In proposing their similarity-coverage model, Osherson et al. (1990) hoped to avoid some of these problems by using fictitious biological properties, about which people have few prior beliefs. Because these predicates are unfamiliar, "they are unlikely to evoke beliefs that cause one argument to have more strength than another" (Osherson, 1990, p. 186), and they can therefore isolate the effects of categories on inductive inference. The results by Gelman and Markman (1986) and Heit and Rubinstein (1994), however, suggest that the biological flavor of these predicates leads people to assume that the predicates will generalize according to specifically biological mechanisms rather than according to overall similarity. In some cases, people may assume that these mechanisms run along the lines of taxonomically related species or superordinate categories. For example, if horses have Vitamin Z, then perhaps whatever internal causal mechanism is responsible for Vitamin Z production in horses is also at work in biologically related species such as donkeys, or even in all mammals; hence, donkeys (or mammals) have Vitamin Z. In other cases, people may reason that the relevant causal mechanism is external to the individual organism. Perhaps

[3] Proponents of the similarity-coverage model could argue that Gelman and Markman (1986) labeled their items in a way that emphasized category membership; hence, the fact that children favored the indirect route in their inferencing was due to features of the experimental setup. For example, Gelman and Markman told children that "this dinosaur" (the brontosaurus) has cold blood, that "this rhinoceros" has warm blood, and asked whether "this dinosaur" (the triceratops) has cold or warm blood. Repetition of "dinosaur" might have emphasized the indirect transmission route through the corresponding category. This argument gains support from Medin, Lynch, Coley, and Atran's (1997) finding that even experts prefer to generalize across natural kinds when their common genus is linguistically marked (e.g., Norway maple and sugar maple) than when it is not (Ohio buckeye and horse chestnut). In Gelman and Markman's experiment, however, the preference for generalization by category cannot be explained by mere repetition of the common noun ("dinosaur"). Gelman and Markman (1986, Experiment 2) found similar results when they used synonyms instead of exact repetitions. Could mention of the category itself (by whatever noun) have driven participants to the indirect route for generalization? Although this is possible, it still would not explain the difference in results for biological and accidental properties.

horses have Vitamin Z by virtue of eating certain kinds of feed, so other animals that eat the same feed should also have Vitamin Z. The similarity-coverage model may provide better fits to the data in the former case, since biological (taxonomic) relationships play a role in determining indirect transmission between species. But in neither case is global similarity likely to be the guiding principle.

In their later work, Osherson, Smith, Shafir, and their colleagues (Osherson, Smith, Myers, Shafir, & Stob, 1994; Osherson, Smith, Shafir, Gualtierotti, & Biolsi, 1995; Smith, Shafir, & Osherson, 1993) have attempted to account for arguments with more familiar properties by proposing a somewhat related theory, called the *gap model.*[4] The model applies to arguments similar to Arguments 1–4 in which the subject nouns of the premises and conclusion name a category. However, the predicates can differ in the premises and conclusion, and they are ones that describe a possible property of category members. Osherson et al.'s examples and experiments typically involve complex predicates, such as *have skins that are more resistant to penetration than most synthetic fibers* or *have a visual system that fully adapts to darkness in less than 5 minutes.* It is possible to illustrate the main features of the gap model, however, using simple one-dimensional predicates, such as *is at least 5 feet tall.* Arguments 5 and 6 are examples of this sort:

(5) Lions are at least 5 feet tall.
———————————
House cats are at least 5 feet tall.

(6) Lions are at least 5 feet tall.
———————————
Cougars are at least 5 feet tall.

There is a discrepancy (a "gap") between the property expressed by the predicate (*are at least 5 feet tall*) and the corresponding property of the members of the premise category (lions). Lions can be 4 feet tall but probably not 5 feet. Evaluating the strength of such arguments requires assuming that the premise is true and then determining to what extent the conclusion is true under that assumption. Assuming the premise to be true, in turn, means reducing or eliminating the gap. According to the gap model, we do this by recalibrating the conclusion category (house cats in Argument 5 and cougars in Argument 6) by adding to it (some function of) the positive difference between the value of the property expressed by the predicate (*is 5 feet tall*) and that possessed by the premise category (lions).

As an illustration of this recalibration process, let's assume that the actual height of lions is 4 feet, cougars 2 feet, and house cats 1 foot. Then the recalibrated size of house cats in the context of Argument 5 will be the actual height of house cats plus the difference between 5 feet and the height of lions: $1 + (5 - 4) = 2$ feet. The gap model assumes, however, that the extent of the adjustment also depends on the similarity between the premise and conclusion categories (and on the similarity between the premise and conclusion predicates, when these differ). For these purposes, similarity is measured on a scale from 0 (*not at all similar*) to 1 (*maximally similar*). This fractional value of similarity multiplies the difference between the value of the predicate and the premise category before adjustment. For example, if the similarity between lions and house cats is .8, then the recalibrated height of house cats will not be 2 feet but $1 + .8(5 - 4) = 1.8$ feet. For the moment, let's assume that the similarity between cougars and lions is the same as that between house cats and lions; if so, cougars' recalibrated height will be $2 + .8(5 - 4) = 2.8$ feet.

To calculate the probability of the argument's conclusion given its premise, the gap model determines how likely it is that the recalibrated conclusion category will have the value of the predicate. The probability of the conclusion of Argument 5, given the premise, is then the probability that 1.8-foot house cats are 5 feet tall, and the probability of the conclusion of Argument 6, given the premise, is the probability that 2.8-foot cougars are 5 feet tall.[5] The gap model computes these probabilities as:

$$Pr(\text{conclusion}|\text{premise}) = \frac{\min(\text{recalibrated value of conclusion category, value of conclusion predicate})}{\text{value of conclusion predicate}}$$

For example, if the recalibrated height of a house cat is 1.8 feet, as above, the probability of the conclusion of Argument 5, given its premise, is minimum(1.8, 5)/5 = .36. Similarly, if the recalibrated height of a cougar is 2.8 feet, then the probability of the conclusion of Argument 6, given its premise, is minimum(2.8, 5)/5 = .56. This accords with the intuition that Argument 6 is stronger than Argument 5.

Questions about the gap model. The examples in Arguments 5 and 6 highlight a peculiarity in the role that similarity plays in the gap model's computations. By contrast with the similarity-coverage model, in which only overall similarity between categories matters, the gap model considers similarity with respect to properties "potentiated by the predicate" (Smith et al., 1993, Footnote 5). This seems on the right track, because, for example, no matter how similar a toy stuffed lion may be to a flesh-and-blood lion, learning that a lion is 5 feet tall may have minimal effects on one's estimate that the toy is 5 feet tall (Carey, 1985). However, how do we determine which properties the predicate potentiates? Restricting the similarity computation to the property that the predicate *denotes* runs into problems. In Arguments 5 and 6, the predicate denotes a value on the dimension of height. Following this possibility would mean determining similarity according to the difference in height between lions and house cats in Argument 5 and between lions and cougars in Argument 6. But why should the recalibrated value of the conclusion category in these arguments depend on the similarity in height between lions and house cats or cougars? Do we really want to adjust the height of cougars more than the height of house cats because the difference in height between cougars and lions is smaller than that between house cats and lions?

Examples like these led Smith et al. (1993, p. 93) to conclude that when the critical dimension is obvious, as height is in Arguments 5 and 6, people no longer use similarity in calculating the

[4] The formulation of the gap model changes from the earlier to the later articles in this series; the description here follows Osherson et al. (1995), as this article introduces improvements and also generalizes the earlier versions.

[5] Of course, taken literally, the probability is 0 that something 1.8 feet tall is 5 feet tall. So we should consider the values to represent approximations or central tendencies rather than exact points. According to this interpretation, the question is how likely it is that something estimated to be 1.8 feet tall could really be 5 feet tall. This difficulty is probably the motivation for Osherson et al.'s (1995) use of more complex predicates whose exact values participants are unlikely to know.

conditional probability of the conclusion. Still, you might well believe that there are some relevant relationships between the premise and the conclusion categories, apart from height, that could affect the strength of arguments like these. Argument 5 becomes less plausible when we substitute for "house cat" an inanimate object of approximately the same size. For example, Argument 7 seems noticeably weaker than Argument 5, even assuming that Coke bottles have exactly the same height as house cats.

(7) Lions are at least 5 feet tall.
Coke bottles are at least 5 feet tall.

It is possible to state this point in a more general way. In evaluating arguments for inductive strength, we are invited to assume that the premises are true, and we must then determine to what extent this assumption changes the believability of the conclusion. In assessing the conclusion, we could take into account the similarity between premise and conclusion categories (as in the similarity-coverage model) and the relative degree to which conclusion and premise categories fall short of the predicate's value (as in the gap model). But we sometimes also need to understand why or how the premise is supposed to be true. This is because the causal factors (or other factors) that would make the premise true may determine how the premise information generalizes (see Heit, 2000, for a similar conclusion). We noticed in discussing the Vitamin Z examples that how the property generalizes might depend on whether we believe the premises are true because the animals produce the vitamin internally or because they obtain it externally from food. Similarly, the results from Lopez et al. (1997) suggest that the believability of the conclusion may depend on people's notions of how an unknown disease is transmitted. And, likewise, the strength of Arguments 5–7 may be a function of how we imagine the premise could become true: If it's a matter of feline growth hormone or some other biological factors, we would probably find Argument 6 stronger than 5, and Argument 5 stronger than 7. If it's a matter of general stretching or other purely physical–mechanical factors, we might find Arguments 5 and 7 equivalently strong (or, more likely, equivalently weak). These considerations of how the premises become true go beyond simple comparison between the categories or between the predicate's properties and those of the categories. They also depend on the causal dependencies that are in place (see Burstein, Collins, & Baker, 1991, and Collins & Michalski, 1989, for a theory of induction based partly on such dependencies).

A pair of studies by Sloman (1994, 1997) has demonstrated the importance of these external causal factors. In these experiments, participants received arguments, such as 8 and 9, and they estimated the conditional probability of the conclusion given the premise. The arguments had premises that suggested an explanation that either carried over to the conclusion (as in Argument 8) or did not carry over (as in Argument 9). For example, ranchers might be required to get rabies vaccines because of their exposure to animals, a risk that is also common to zoologists. Ranchers try to control animal breeding, however, in order to improve their livestock, a goal that may not be relevant to zoologists.

(8) Ranchers are required by law to receive rabies vaccines regularly.
Zoologists are required by law to receive rabies vaccines regularly.

(9) Ranchers try to control the breeding of animals.
Zoologists try to control the breeding of animals.

Sloman found that participants' estimates of the conditional probability of the same-explanation items was higher than their estimates of the probability of the conclusion alone. The premise boosted the likelihood of the conclusion for these arguments. By contrast, the premise of the different-explanation items either reduced the conditional probability of the conclusion or produced the same estimate as the isolated conclusion. To account for these data, a proponent of the gap model would have to contend that (a) the predicate of these arguments evokes a corresponding set of properties in the representation of the categories (animal handling frequency? frequency of controlling breeding?), (b) the conclusion category's values on these dimensions are recalibrated, and (c) the probability of zoologists having the original property is assessed relative to the recalibrated values. But granting these assumptions, it is unclear why the conclusion probability of Argument 9 would remain unchanged or decrease. A simpler hypothesis is that the causes for requiring rabies vaccinations apply to both ranchers and zoologists, whereas the causes for wanting to control breeding of animals apply to ranchers but are irrelevant to zoologists.

Hypothesis-testing theories of category-based induction. The idea that general causal knowledge affects the inductive strength of arguments suggests that we could treat the conclusion of arguments, such as 1–9, as analogous to a scientific hypothesis and the premises of the argument as evidence for this hypothesis. We could then use theories of hypothesis testing or confirmation to explain judgments of inductive strength as a special case. This idea goes back at least to Carnap (1950). Recently, Heit (1998, 2000) and McDonald, Samuels, and Rispoli (1996) have proposed psychological models of category-based induction along these lines. For Heit, the probability of the conclusion given the premises is just the probability of a hypothesis (that the conclusion is true) given the evidence (provided by the premises). So we can apply Bayes's theorem to obtain this probability, assuming some prior distribution of probability over potential hypotheses. For example, in the case of Argument 10, if we have the prior probability that a hypothesis will be true of cows, horses, mice, and all other mammals, and the prior probability that it will be true of cows, horses, and mice, but not all other mammals, then Bayes's theorem allows us to calculate the probability that the hypothesis is true of all mammals, *given* that it is true of cows, horses, and mice (see also Tenenbaum & Griffiths, 2001).

(10) Cows have Vitamin Z.
Horses have Vitamin Z.
Mice have Vitamin Z.
Mammals have Vitamin Z.

McDonald et al. (1996) proposed that factors that affect hypothesis testing in empirical studies—in particular, the amount of evidence, the scope or range of the given hypothesis or conclusion, and the number of alternative hypotheses—can also predict judged argument strength. In the case of arguments like 10, the amount of evidence comes down to the number of (mammal) subcategories that the premises specify as having the property in question (cows,

horses, and mice, in this example),[6] the scope of the hypothesis is the size of the conclusion category (the total number of mammals), and the alternative hypotheses are possible alternative explanations that the premises bring to mind (e.g., that having Vitamin Z is restricted to land-based mammals). Heit (1998) and McDonald et al. both showed that hypothesis-testing theories can account for many of the same phenomena that the similarity-coverage model does. McDonald et al. also produced impressive correlations to actual judgments of argument strength.

These hypothesis-testing theories, unlike the similarity-coverage approach, also generalize immediately to arguments about familiar properties. For example, if we have the prior probabilities that a specific property (e.g., having good eyesight) is true of the different combinations of the premise and conclusion categories, we can again plug into Bayes's theorem to compute the conditional probability that the property holds true of the conclusion category given that it holds true of the premises' categories. It's controversial, of course, whether people use Bayes's theorem to test hypotheses (see, e.g., A. Tversky & Kahneman, 1974), and Heit (1998) proposes his theory as a theoretical account of the goals of inductive inference (what needs to be computed) rather than as a processing model. In addition, the ability of the model to describe or explain phenomena from experiments on induction is a function of the prior distributions, and as Heit notes, Bayes's theorem provides no account of where these distributions come from. Heit plausibly suggests that people might estimate the distributions from known properties of the categories in question or from higher order beliefs about the distributions of properties across categories (Shipley, 1993). The difficulty, then, is similar to one we met in assessing the gap model: To make the gap theory work properly, we need to know the underlying properties responsible for the "gap"; to make the Bayesian model work properly, we need to know the underlying processes that determine property distributions. This may itself require an additional kind of reasoning not specified by the theory, as Heit (1998) acknowledges.

The role that prior probabilities play in Heit's (1998) approach corresponds in part to the number of alternative hypotheses in McDonald et al.'s (1996). In their experiments, McDonald et al. measure the alternatives empirically by providing participants with a list of premises from arguments like 10 and asking them to construct hypotheses about objects that might reasonably have the property. This measure significantly predicted judged argument strength from a separate group of participants who inspected the full arguments (same premises plus conclusions). The success of this prediction, though, raises the issue of how the participants arrived at their alternative hypotheses. What about the premises tempted participants to suppose that the property in question might generalize in one way rather than another? Because this is equivalent to the question of how people make inductive inferences, this version of the hypothesis-testing theory also presupposes some important reasoning processes that occur off stage. These theories provide a general and useful framework for thinking about argument strength but leave unexplored some cognitive prerequisites that are essential for the theories' success.

Summary

What does category-based inductive inference tell us about the nature of categories? Natural categories and their properties are not uniformly scattered in a vast property soup, but they cluster in ways that support further inferences even about unfamiliar properties. To take advantage of this nonuniformity and to project properties across categories, people may reason that similar categories support similar properties, either directly (Sloman, 1993), taxonomically, or both (Osherson et al., 1990; Rips, 1975). Or they may reason more abstractly that the distribution of new properties should follow the distribution of old ones (Heit, 1998) or follow the contours of linguistic practices in naming (Coley, Medin, & Atran, 1998). What hooks natural categories to their properties, however, are often causal laws, and it would be surprising if people weren't able, at least on some occasions, to use these laws or their instantiations to support inferences. Lassaline (1996) provided evidence that when people have explicit causal information connecting known properties to properties they are trying to generalize, this information increases judged argument strength without increasing judged similarity between the premise and conclusion categories. Rehder and Hastie (2001) also found that people generalize more from instances that embody known causal relations within a category than from instances that violate one or more causal relations.

The gap model (Osherson et al., 1994, 1995; Smith et al., 1993) captures an essential insight in postulating that people conceive ways in which the inductive premises could become true, adjust the conclusion category in light of these alterations, and evaluate the strength of the argument as the likelihood of the conclusion in this changed context. In outline, this idea resembles proposed methods for evaluating the truth value of counterfactual conditionals, such as *if lions were 5 feet tall, then house cats would be 2 feet* (see, e.g., Levi, 1996; Stalnaker, 1968). According to this method, you judge the counterfactual by revising your beliefs to accommodate the antecedent (lions are 5 feet tall). If the consequent information (house cats are 2 feet tall) is true in the revised set of beliefs, then the counterfactual as a whole is true as well. This correspondence between assessing the inductive strength of arguments and assessing the truth of counterfactuals is not surprising considering that the same causal principles may sometimes support both of them (Goodman, 1955). Current psychological models of category-based induction, however, scant the details about how people carry out the belief adjustment that is central to this endeavor. I have tried to argue that the process isn't necessarily as simple as revising values on prespecified dimensions. Instead, we use our knowledge of which aspects of categories are changeable, what the causes of these changes are likely to be, and how the consequences of these changes affect other categories. This is not intended as a theory of category-based induction, but it may point to ingredients missing in current theories. Perhaps the best way of

[6] Sheer number of subcategories might not be the best measure of amount of evidence, as McDonald et al. (1996) acknowledge. For example, replacing cows with bison, horses with zebras, and mice with voles seems to make Argument 10 weaker, presumably because the new categories are less frequent or less important members of the mammal category. It's unclear from McDonald et al.'s discussion how amount of evidence is best analyzed.

viewing current research in this area is as illustrating default strategies that people adopt when more explicit knowledge is unavailable.

Natural Categories and Their Transformations

The research that we have just examined derives information about categories from the role they play in inductive reasoning. Positing new information about a category in the premise of an argument can force us to modify our beliefs about the category for purposes of the inference. Which modifications we make—which aspects of the category are easily modifiable and which are not—can provide evidence about (our beliefs about) the category's structure. (Nisbett, Krantz, Jepson, & Kunda, 1983, made a similar point in terms of category–property homogeneity.) Subjecting a category to this sort of inferential pressure gives us a test of the category's makeup. Because the properties that the premises ascribe to the category can be counterfactual, the induction paradigm identifies beliefs about what *might* be true or what *must* be true of the category in unrealized situations. These are beliefs about a category's modal properties, not merely beliefs about what is normal or average in our own experiences.

There is, however, another way to examine these modal properties. Instead of attributing a property to a category (or to an individual category member) and studying how the property generalizes to others, you can change a property of a member and check whether the individual retains its category membership. Consider, for example, beliefs about Leigh, an individual lion. Changes in Leigh's external appearance are perfectly consistent with her remaining a lion, whereas other changes, particularly in her internal makeup, cause both children (Gelman & Wellman, 1991; Keil, 1989) and adults (Barton & Komatsu, 1989; Rips, 1989) to think that Leigh no longer counts as a lion. These judgments, of course, rely on inductive inference, just as judgments about explicit arguments do in the research discussed earlier. The corresponding argument here might be similar to Argument 11:

(11) Leigh is a lion at time t.
Leigh undergoes cosmetic surgery so that her external appearance becomes identical to that of a tiger at $t + 1$.

Leigh is a lion at $t + 1$.

However, because the nature of such arguments differ in content from those in experiments on category-based induction, I consider these inferences separately here.

Evidence From Transformations of Category Members

Keil's (1989) studies of natural kinds created conflicts between the appearance of an individual organism and the more fundamental properties of its inner constitution, parentage, or progeny. Some of his experiments informed children about discoveries in which scientists find that an organism that appeared to be a member of one category (e.g., horses) has the inner parts, parents, and offspring of another (e.g., cows). Other experiments provided stories of normal organisms of one category whose external appearance changes permanently to resemble that of another category (e.g., a horse that a doctor alters to have stripes and to eat wild grasses like a zebra). Both sets of studies provided evidence that between kindergarten and second grade, children come to appreciate the more theoretically important properties and to discount the more superficial ones. Further growth in this knowledge continues through at least fourth grade. Keil (1989) argued, however, against the view that kindergartners are prisoners of external appearance (see also Keil, Smith, Simons, & Levin, 1998). Temporary changes (e.g., a horse in zebra costume or a horse with stripes that wash off in the rain) do not lead kindergartners to suppose that an organism has switched categories. Nor do changes that make an animal resemble an inanimate object (e.g., a porcupine made to look like a cactus) convince them that the animal has transmuted.

In simple settings, even preschool children are sensitive to the importance of internal properties (Gelman & Wellman, 1991). Four- and five-year-olds usually deny that an animal whose insides have been removed is still an animal. For example, they answer "no" to grizzly questions like, "What if you take out the stuff inside of the dog, you know, the blood and bones and things like that and got rid of it and all you have left are the outsides? Is it still a dog?" At the same time, they usually affirm that an animal whose outsides are removed is still an animal. They answer "yes" to, "What if you take off the stuff outside of the dog, you know, the fur and got rid of it and all you have left are the insides? Is it still a dog?" They also know that natural kinds are likely to have natural-kind insides, whereas artifacts have artifact insides, despite lack of detailed knowledge about the insides' structure (Simons & Keil, 1995). Moreover, older four-year-olds appreciate that an animal or plant of one species raised among those of another species will retain its category membership—for example, that a watermelon seed planted in a corn field will produce watermelons rather than corn (Gelman & Wellman, 1991). It is less clear, however, that preschool children are able to predict correctly which of an organism's properties—for example, its physical traits versus its beliefs and preferences—are the likely products of its birth parents and which are the products of its adoptive parents (Solomon, Johnson, Zaitchik, & Carey, 1996; Springer, 1996).[7]

Controversy about these results (and the developmental results we glimpsed earlier) centers on the question of whether they reflect increasing sophistication of a preexisting base of biological knowledge or, instead, the emergence of biological knowledge from a nonbiological—social or psychological—precursor (e.g., Atran, 1998; Carey, 1985, 1995; Inagaki & Hatano, 1993; Johnson & Carey, 1998; Keil, 1995). For our purposes, however, the outcome of this controversy isn't as important as the clues the experiments yield about people's eventual beliefs about natural kinds. Because children have never witnessed horses cross-

[7] A potential ambiguity about these results is that they may reflect children's pragmatic uncertainties about labeling rather than their beliefs about category membership. They may believe, for example, that a dog whose fur has been shaved is no longer a dog but still have no better word for it than "dog." The yes–no format of the questions may reduce this worry, since children don't have to produce their own label for the furless creature, but it is still possible that when children are asked "Is it still a dog?" they hesitate to say "no" for lack of a better descriptor. This alternative view must explain, however, why children do say "no" when they are told that the dog's insides are removed. This view would be forced to the position that both transformations cause children to believe that the creature is a nondog but that only the more radical transformation is enough to overcome the tendency to agree to the "dog" label.

dressing as zebras or horses that have undergone cosmetic surgery to look like zebras, their answers don't reflect mere knowledge of these events. Analogies from transformations that children *have* witnessed might be a source of information, but if this is so, the analogies must take into account the fact that transformations preserve category membership in some domains but not others. Older children realize that changing the external appearance of a horse can't change it to a zebra, but changing the external appearance of a coffee pot may well change it into a bird feeder (Keil, 1989). They also know that internally caused changes in size and parts are permissible for animals but not for artifacts, such as lightbulbs or telephones (Hall, 1998; Rosengren, Gelman, Kalish, & McCormick, 1991). By the time they are adults, people's analogizing, if any, is probably not based on pure similarity, since it is possible to show a double dissociation between judgments of similarity and judgments of category membership for these transformations. In one experiment (Rips, 1989), participants read stories about a member of one natural kind (e.g., a reptile) who undergoes a transformation to resemble a member of another kind (e.g., a fish) but is still able to have normal offspring of the first kind. These participants rated the transformed animal more likely to be a member of the first kind but more similar to the second. In a separate study, participants read stories about animals whose immature form resembles one category but whose mature form resembles another. Participants rated the immature form as more likely to be a member of the second kind but as more similar to the first.

It seems reasonable to suppose, then, that older children and adults possess relatively abstract knowledge that certain sorts of properties are important to category membership, that other properties are not as important, and that which properties are which depends on the domain of the object in question (Barton & Komatsu, 1989; Keil, 1995). For example, most adults judge that molecular structure (and not external appearance) determines which individuals are members of animal and plant categories. Molecular structure, however, is clearly less important for artifact categories than for natural kinds (Barton & Komatsu, 1989).

Essentialist Interpretations of the Transformation Studies

Do the results just discussed show more than that some properties are more important than others for category membership? Do they imply that people hold some properties of objects to be essential for category membership? In examining this issue, we can begin with a recent formulation by Gelman and Hirschfeld (1999), as they have taken pains to clarify the scope of essentialist ideas. First, Gelman and Hirschfeld distinguished their position from earlier philosophical views: Essentialism in this context is a psychological claim about people's beliefs—beliefs about the makeup of natural kinds and certain other categories—not a claim about the actual (metaphysical) composition of these kinds (see also Medin, 1989; Medin & Ortony, 1989). Second, Gelman and Hirschfeld distinguished the *causal* essentialism they promote from a *sortal* essentialism that deals with word meaning. Causal essentialism is belief in a "substance, power, quality, process, relationship, or entity that *causes* other category-typical properties to emerge and be sustained and confers identity," whereas sortal essentialism is knowledge of a "set of defining characteristics that all and only members of a category have" (Gelman & Hirschfeld, 1999, pp. 405–406). Gelman and Hirschfeld rejected sortal essences on the grounds that "given the past thirty years of research on categorization, it is extremely unlikely that people represent features that can identify all and only members of a category . . . regardless of how confident they are that such features exist" (p. 407; see the section *Natural Categories and Their Definitions,* below, for further discussion of this claim).

In Table 1 I attempt to flesh out the claims of causal essentialism in a way that is consistent with psychological views of this topic. Causal essentialism is a theory about people's everyday beliefs about natural categories, and so the characteristics in the table have the status of beliefs. Thus, causal essentialism holds at a minimum that people believe essential forces are responsible for particular objects being members of natural kinds and for the typical properties that these objects have as members. Table 1 displays these two characteristics of essentialism under the headings *potency* and

Table 1
Possible Characteristics of Cognitive Essentialism About Natural Kinds

Characteristic	Description
Potency	Essential properties are responsible for an object being a member of a natural kind.
Productivity	Essential properties are responsible for (a possibly unlimited number of) a member's other properties.
Objectivity	Essential properties exist in nature (do not depend on human convention).
Intrinsicness	Essential properties exist within individual category members (do not depend on other objects).
Uniqueness	Natural kinds have one (or, at most, a small subset of) essential properties common to all members.
Distinctiveness	Different natural kinds have different essential properties.
Identity of members	Essential properties are responsible for tracing the same member of the kind across possible situations.
Identity of individuals?	Essential properties are responsible for tracing the same individual across possible situations.
Discreteness?	An object has the essential properties of a natural kind either completely or not at all.
Prepotency?	No additional factors can override essential properties.

Note. As part of a psychological theory, the descriptions should be prefaced by "People believe that . . . ." Question marks indicate characteristics considered optional.

productivity. In addition, people believe these essential forces are *objective,* existing in nature apart from people's interests and beliefs. However, if causal essentialism were just a belief that something or other (some natural "substance, power, quality," as Gelman & Hirschfeld, 1999, described it) causes the properties of category members, it would be unobjectionable but toothless. It's obvious that people believe that something causes lions to have the properties they possess. A more interesting version of the theory would be that people not only believe that there exist such causes but can actually describe these causes. Part of the doctrine of psychological essentialism, however, is that people often are unable to describe such causes in any detail, sometimes representing them simply as a wild card or "placeholder" (Medin & Ortony, 1989).

A second possible strengthening of causal essentialism that is closer to Gelman and Hirschfeld's (1999) is that causal essentialists believe not just that the properties of category members are caused but that the same cause is responsible for all the typical properties of all members of a category. This cause is *intrinsic,* subsisting in the individual members and independent of other objects. The essential cause is also a *unique* cause that is responsible for all Leigh's liony properties, and for other lions' liony properties as well. Thus, the essential properties provide a unitary explanation for what are otherwise merely correlated external traits. Presumably, also, *distinct* causes produce the typical properties of other categories, so that essential causes differentiate the categories. This claim about belief in unique and distinctive causes is an interesting one, as it is possible that many of the categories that our animal and plant terms denote are not in fact associated with such causes (see, e.g., Dupré, 1993; Sober, 1980; and the discussion in the second part of this article). In fact, however, it is not easy to be precise about how uniqueness and distinctiveness play out in causal essentialism. In the case of uniqueness, for example, it seems consistent with the spirit of the proposal that a small number of causal factors might be jointly responsible for Leigh's lionhood. However, the possibility of a large number of alternative causes does seem incompatible with essentialist intuitions. Perhaps uniqueness and distinctiveness should be spelled out in terms of belief in individually necessary and jointly sufficient causal factors, but this reformulation may also be unclear for reasons discussed later (see *Natural Categories and Their Definitions*). It seems possible (even likely) that people's beliefs are themselves imprecise, going little beyond the notion that lions have a root cause and tigers another.

In their definition of causal essentialism, quoted above, Gelman and Hirschfeld (1999) also asserted that causal essences confer "identity" on category members. This could mean that the causes are responsible for an individual being a category member—for example, for Leigh's identity as a lion in good standing. This is the characteristic that we have already labeled *potency* in Table 1, and the studies just cited bear on this claim. However, "identity" in this context could also mean the object's continued existence as the numerically same member (or even as the numerically same individual) across situations. As another Michigan essentialist put it,

> People in diverse cultures consider . . . essence responsible for the organism's identity as a complex, self-preserving entity governed by dynamic internal processes that are lawful even when hidden. This hidden essence maintains the organism's integrity even as it causes the organism to grow, change form, and reproduce. (Atran, 1998, p. 548)

In this sense, essential causes are responsible not only for Leigh being a lion but also for her being the very same lion (or same individual) in different settings and at different times. It is possible to examine this idea by giving participants stories about transformations that Leigh undergoes and asking whether the transformed organism is still the same lion or is still Leigh (rather than whether she is still a lion; see Blok, Newman, Behr, & Rips, 2001; Hall, 1998; Liittschwager, 1994).[8] For example, Hall (1998) showed children and adults a series of photos depicting a novel object that loses each of its parts one at a time and has these parts replaced with new ones at each step. Subsequently, a person reassembled the old parts into a similar whole. Participants then had to choose either the object with new parts or the object with old parts as the one identical to the original. (This task is based on the philosophical puzzle about the ship of *Theseus* in Hobbes, 1839–1845, part 2, chap. 11.) Adults and 7-year-olds (but usually not 5-year-olds) chose the object with new parts as the same as the original when (a) the object was described as an animal and (b) no human intervention caused the loss of parts. In other conditions (where the object was described as an artifact or where humans intervened in detaching the parts), participants judged the reassembled object the same or divided their votes between the two candidates.

We can take the causal essentialist doctrine to mean that a unique essence is causally responsible for each individual lion's membership in the lion category, for its lionlike properties, and for its identity as the same lion or the same object in different possible situations; a distinct essence is responsible for each individual tiger's tigerlike properties, for its membership in the tiger category; and so on. Table 1 summarizes these characteristics of psychological essentialism, along with some other potential characteristics to be considered later. Because there is ambiguity about whether the essential properties supply criteria of identity for members of natural kinds as such or identity for individual objects in themselves, I distinguish these characteristics in the table, labeling the first *identity of members* and the second *identity of individuals.* Psychological essentialists seem committed to at least the first of these traits (e.g., identity as the same lion). Commitment to the second (e.g., identity as Leigh) is not so clear, and I register it in the table as an optional characteristic of essentialist doctrine. (See Footnote 8 and the section *The Intrinsic View,* below, for further discussion of identity.)

The characteristics of Table 1 are important here because essential properties are one obvious source of natural kinds' modal qualities. The identity characteristic, in particular, permits a way of thinking about which object is the same lion in different possible situations and so a notion of the range of properties that could possibly be true of her. It is therefore important to examine

[8] The distinctions among "is still a lion," "is still the same lion," and "is still Leigh" are subtle but important in thinking about natural kind's modal properties. An organism could still be Leigh without still being a lion after a transformation if it is possible for Leigh to be a member of another species in some possible worlds. Similarly, an organism could still be a lion without still being the same lion if it is possible for lions to trade identities in some possible worlds. Which of these distinctions people observe is an open question at this time.

essentialist theories closely. An immediate question, then, is whether the evidence supports causal essentialism. Do people think natural kinds possess unique and distinctive causal essences or do they merely hold a minimal view that there are some causes or other that natural kinds have that are responsible for their properties, membership, and persistence (Strevens, 2000)?

Questions About the Transformation Studies

The studies I have reviewed provide evidence about which properties of objects sustain membership in a category across possible transformations. Internal mechanisms and descent, for example, are important in this respect, whereas external appearance and location are not. The studies may also support the idea that people take the former factors as essential for category membership, in some sense of "essential" that we have begun to fill out. Most of the criticism of these experiments has focused on essentialism, and I examine these issues here as they come up in recent antiessentialist experiments. (Discussion of psychological essentialism's theoretical pros and cons is deferred to the second part of this article.)

A more general objection to the transformation studies, however, is that although they tap people's higher level thinking about natural categories, they shed no light on how people recognize category members in everyday encounters (Smith & Sloman, 1994). In deciding whether an animal you are observing is a horse or a zebra, you don't typically examine the animal's pedigree or genetic markers but instead rely on superficial perceptual properties in making the decision. You may use deeper—unobservable or theoretical—aspects of the organism mainly in special situations when this information is at hand, when correct classification or inference is important (for scientific purposes, say), or when no superficial properties happen to be available. This point is well taken, and it limits the scope of conclusions from the transformation studies. The purpose here, however, is to examine people's beliefs about natural categories' modal properties, and for this reason we must look beyond perceptual recognition and categorizing, as noted at the outset.

Objections to specifically essentialist claims have focused on two issues: one having to do with the relation between essence and membership, and the other with transformations of members versus entire species. The thinking behind the first of these problems is that if essences are unique and distinctive in the way we supposed earlier, then something is a horse if and only if it has horse essence. There should be no intermediate cases of animals that are only partly horses. However, Kalish (1995, Experiment 1) found that participants rate atypical organisms as members of natural kinds to some degree; for example, they rated a zebra as "sort of a horse" and a wolf "sort of a dog." As Kalish notes, these intermediate ratings might reflect the uncertainty of participants' beliefs about category membership (e.g., McCloskey & Glucksberg, 1978) rather than their belief that category membership is uncertain: They may be unsure whether zebras are horses rather than being sure that zebras are partial horses. In addition, it might be possible for an essentialist to suppose (as do Gelman & Hirschfeld, 1999) that an organism can possess essential properties to a greater or lesser extent. If so, then essentialism is compatible with categories that are graded rather than all-or-none, in accord with Kalish's data.[9]

This objection isn't quite so easily evaded, though. Even if essentialists allow objects to possess an essential property to a variable degree (so that, for example, a zebra can have a partial helping of horse essence), we would at least expect degree of membership to track degree of essence. The more essence of a natural kind something has, the better a member of the kind it should be. For instance, assuming that H_2O is the essence of water, then the more H_2O a substance has, the better it should be as a type of water. Malt (1994) has shown, however, that whether people call a substance water is not even monotonically related to their belief about the percentage of H_2O in the substance. For example, participants judged ocean water to contain 79% H_2O but saliva (a nonwater) to contain 89%. (Not all of Malt's examples of waters and nonwaters involve natural kinds. For example, the water categories included radiator water and sewer water, which may be human artifacts. But the overlap in percentage of H_2O persists even for more natural substances, as in the above examples.)

The second type of experimental objection to psychological essentialism comes from judgments about further discoveries and transformations. Braisby, Franks, and Hampton (1996) provided evidence that discoveries about the intrinsic properties of natural categories do not always affect participants' judgments about the existence of these categories. In the key conditions in this experiment, participants read stories such as 12a, in which an individual category member is discovered to lack a key property of the category, and stories such as 12b, in which all members of the category are discovered to lack the property (Braisby et al., 1996, p. 256):

(12) a. You have a female pet cat named Tibby who has been rather unwell of late. Although cats are known to be mammals, the vet, on examining Tibby carefully, finds that she is, in fact, a robot controlled from Mars.

b. You have a female pet cat named Tibby. For many years people assumed cats to be mammals. However, scientists have recently discovered that they are *all,* in fact, robots controlled from Mars. Upon close examination, you discover that Tibby too is a robot, just as the scientists suggest.

Participants then answered questions about the existence of cats and about whether Tibby is a cat, given each discovery.

The predictions that these authors make on behalf of essentialism are based on the philosophical theories of reference developed by Kripke (1972) and Putnam (1975). These theories hold, roughly speaking, that the referent of natural-kind expressions, such as "water" or "cat," is fixed at the time of their introduction by local samples of the kind in question. Thus, whether an arbitrary specimen falls under these terms depends on whether it is in the same kind as that of the local sample. If present-day scientific theories are correct, then whether a substance is correctly termed "water"

[9] Some versions of essentialism, however, do include the idea of all-or-none categories; see Ellis (1996) for one such version. In further research Diesendruck and Gelman (1999) found a greater number of all-or-none judgments for animal categories than for artifact categories (i.e., participants were more likely to say that something was either definitely a fish or definitely not a fish than that something was either definitely a tool or definitely not a tool). Even for animal categories, however, there were some intermediate membership judgments. See also Malt (1990).

depends on whether its molecular structure is identical to that of the local samples, and whether an object is correctly termed a "cat" depends on whether its genetic structure (or other underlying properties or relations) is the same as that of the original cat examples.[10] In the case of stories like 12a, if participants (a) share the Putnam–Kripke intuitions and (b) interpret the story to mean that Tibby is discovered not to possess the property that determines the same-kind relation to original cat samples, then they should judge that the kind cat exists but that Tibby is not a member of the kind. For story 12b, if participants assume that the same-kind relation for cats now depends on being a Martian-controlled robot, then they should again assert that cats exist but, this time, that Tibby remains a cat (see Braisby et al., 1996, Table 1). Braisby et al. found 47–89% agreement with essentialist predictions for the 12a-type stories and 73–87% agreement with these predictions for the 12b-type stories (where the range depended on how the questions were framed; see Braisby et al., 1996, Table 7). These investigators concluded that "our evidence indicates that people do not, in fact, believe that things have essences, if essences are interpreted according to the model provided by Kripke and Putnam (even though people may sometimes behave as if they did)" (Braisby et al., 1996, p. 270).

One reaction to both Malt's (1994) and Braisby et al.'s (1996) findings is that they are irrelevant to the claims about causal essentialism that we glimpsed in the preceding section. Gelman and Hirschfeld (1999) complained on this score that "critically, H_2O represents a sortal not causal essence, and accordingly [Malt's] study provides evidence only against a classical view of category meaning" (p. 408). Similarly, they claimed against Braisby et al. that "on a *causal* essentialist view, the essence need not provide necessary and sufficient clues for determining reference . . . and accordingly the experiments are relevant to a sortal (not causal) essentialist view" (p. 408).

Psychological essentialists and antiessentialists agree that people do not possess a set of necessary and sufficient criteria that determine the meanings of the terms they use. Essentialists' rejection of "sortal essentialism" secures this agreement. The issue that divides these groups is therefore whether the antiessentialists' experiments cast doubt on belief in essence of a more abstract sort. Gelman and Hirschfeld's (1999) position may be that these results do no more than provide further evidence that people don't know what the essential (necessary-and-sufficient) features are, but antiessentialists might well contend that this is too narrow an interpretation. The intent of Braisby et al. (1996) was not merely to show that people lack *mammal* as a necessary feature for *cat* but to prove that they also lack the higher-order belief that essence determines the denotation of natural kind terms (see the quotation from Braisby et al. in the paragraph before last). The claims on both sides are metacognitive ones: whether beliefs about essence play an important role in thinking about kinds.

One way to reconcile essentialists and antiessentialists would be to suppose that essentialists are right about people's theories of kind's physical makeup, whereas antiessentialists are right about people's theories of meaning. Perhaps people think that essences cause animals, plants, and other natural categories to have the properties they do, but they don't believe that essences play a role in determining the referents of expressions for these kinds. This would be to invoke the distinction between causal and sortal essentialism at a higher level: People's ideas about reference and meaning could be partly independent from their ideas about biology, chemistry, physics, and other domains of natural kinds.

There may be limits, however, to the distance that can separate beliefs about the nature of kinds from beliefs about the meaning of kind terms—between ideas of what a kind is and of what the term for the kind applies to. Suppose you are a causal essentialist and believe that there are causal factors distinguishing lions from other species. It would be odd if the news that these causal factors were absent in Leigh did not affect your belief about whether the term *lion* correctly refers to her. Thus, it is not clear how much room causal essentialists have to maneuver between (a) object *O* has the causally essential properties of a category, *C*, and (b) the name of *C* correctly refers to *O*. If a property is causally crucial in determining whether something is a lion, then it is also crucial in determining whether "lion" is true of it in causally possible circumstances. To put this slightly differently: What reason could there be for affirming that people believe natural categories have essences while denying that they believe that terms for natural categories are associated with necessary and sufficient properties? There may indeed be uncertainties about what sorts of properties can function as necessary and sufficient for purposes of meaning, as I discuss later, but it is hard to see why such qualms wouldn't apply equally to essences.[11]

Recall, too, that the evidence in favor of causal essentialism that was reviewed earlier depends on altering causal factors that surround a member of a kind and quizzing participants about whether it remains a category member—for example, whether a lion whose insides are scooped out is "still a lion" (Gelman & Wellman, 1991) or whether a goat with altered chromosomes is "still a goat" (Barton & Komatsu, 1989). This type of question is, in fact, not very different from that posed in Story 12a, so denying the rele-

---

[10] According to these theories of reference, which properties are essential depends ultimately on which properties actually do determine sameness of kind to local samples, not on what current scientific theories happen to say. So whether H_2O is an essential property of water depends on whether present-day chemistry is true. In more recent work, Putnam (1990) allows for greater distance between everyday use of natural kind terms and their use in science:

> I would distinguish ordinary questions of substance-identity from scientific questions. I still believe that ordinary language and scientific language are interdependent; but layman's "water" is not the chemically pure water of the scientist, and just what "impurities" make something no longer water but something else (say, 'coffee') is not determined by scientific theory. (p. 69)

Malt's (1994) evidence on everyday uses of "water" seems to confirm this lay sense of the term. See also Boyd's (1999) distinction between everyday natural kinds and scientific natural kinds.

[11] This is not to say that properties that are central to membership in a natural category are exactly the same as those that people use to pick out members of the kind or those they believe are most important in applying the kind term (Sloman & Ahn, 1999). Whether an animal has a mane may be important in whether *lion* appropriately applies to it, but having a mane is not causally crucial for determining whether it is a lion. This is because a term's appropriateness is partly a pragmatic matter; it depends on not misleading others. Nevertheless, the point remains that if essence determines kind membership, then essence also determines whether it's correct to apply the name for the kind to an instance.

vance of Malt's and Braisby et al.'s studies may be self-defeating for causal essentialists.

The results of Braisby et al. (1996) and Malt (1994) bear on causal essentialism as a psychological theory, but it is not clear that they defeat it. A majority of participants supported essentialist predictions in Braisby et al.'s experiments, so the difficulties these data pose for essentialism depend on how seriously one views departures from complete agreement. As just noted, a causal essentialist could dissent from some of the Kripke–Putnam intuitions about Stories 12a and 12b without sacrificing the idea that essential underlying causes determine kind status. Such an individual might believe, for example, that what fixes cathood once and for all is having a brain of a certain sort and not other causes that scientists happen to discover. Such a person would judge that there are no cats in worlds in which all catlike objects are robots, such as that in Story 12b. This would cause the person to depart from Braisby et al.'s essentialist predictions, although the person could still be said to be a causal (and even a sortal) essentialist.[12] In the case of Malt's (1994) experiment, essentialists might invoke the sorts of pragmatic considerations mentioned in Footnote 11 (see Abbott, 1997, for an argument of this sort).

Some difficult issues remain, though. I have already raised the question of whether possession of essences must be all or none. A second question is whether it is consistent for an essentialist to think that factors in addition to underlying causal ones could also affect what is water, factors like the use to which the substance is put or the location in which it is found (Malt, 1994; see also Hampton, 1995b). Similarly, could an essentialist believe, for example, that H_2O determines what is water but also that certain impurities (e.g., tea extract) disqualify a mixture as water whereas others (e.g., soil) do not? To what extent can causal essentialism admit exceptions in causally possible circumstances? These issues depend on further details of the essentialist position, and we postpone discussion until we have had a look at some additional evidence. To record uncertainty about these matters for now, Table 1 lists the characteristics of *discreteness* (essences are all or none) and *prepotency* (nothing can override essences) with question marks to indicate that these items are ones on which essentialists might disagree.

Direct Assessments of Causal Structure

The transformation experiments suggest that 4–7-year-olds come to think underlying causal properties are important in deciding membership in natural kinds, more important than properties of members' external appearance. It would therefore be sensible to expect that the stronger or more central a causal factor is—for example, the more effects it has—the more important for membership it might be. In line with this prediction, Ahn (1998, Experiment 1) reported a negative correlation between (adult) participants' ratings of the likelihood that a specific factor causes other properties for a given kind and their ratings that members of the kind could lack that factor. For example, participants judged that having goat genetic code was very likely to cause a goat to give milk and to have four legs, and they also judged that it was very unlikely that "a goat would still be a goat if it were in all ways like a goat except that it did not have a goat's genetic code."

It is also possible to explore the role of causal factors by constructing artificial "natural" kinds in which these factors explicitly vary (Ahn, 1998; Rehder & Hastie, 2001; Sloman, Love, & Ahn, 1998). Participants might learn, for example, about a fictitious type of flower that has certain attributes, with causal (or other) relations that run between these attributes. The participants then judge whether novel instances that possess some of these attributes and lack others belong to the category. By varying the attributes and their relations, investigators have used this technique to determine whether causal status of an attribute (central vs. peripheral cause), causal structure of the kind (one cause with many effects vs. many causes of a single effect), qualitative nature of the characteristic (molecular vs. functional), and type of kind (natural vs. artifact) affect category decisions. The results from these studies suggest that participants' knowledge about the relations between attributes is critical for category membership; that lack of an interattribute relation can cause participants to decide that an instance is unlikely to be a category member (Ahn, 1998; Rehder & Hastie, 2001; Sloman et al., 1998). In one study, for example, participants learned that phyrum flowers have velvety leaves that repel mosquitoes; they were then more likely to class as phyra velvety-leaved, mosquito-repelling flowers and flowers lacking both these properties than flowers having one property but not the other (Rehder & Hastie, 2001).

Most other findings with this paradigm have been negative or conflicting, however. There is no evidence that internal attributes (e.g., having eucalyptol in their petals) matter more for membership in natural kinds than do functional attributes (attracting insects) when causal status is constant, and there is no evidence that internal attributes are more important for natural kinds than for

[12] Other departures from essentialist predictions in Braisby et al. (1996) may depend on details of the wording of the stories and the probe questions. In the case of stories like 12b, in which a discovery is made about all cats, Braisby et al.'s results show that on 87% of trials participants endorsed Statement A and rejected B, in accord with the essentialist predictions (see their Table 7):

A. Cats exist.
B. Cats do not exist.
C. Cats do exist and people's beliefs concerning cats have changed.
D. There are no such things as cats, only robots controlled from Mars.
E. Tibby is a cat, though we were wrong about her being a mammal.
F. Tibby is not a cat, though she is a robot controlled from Mars.

Most participants also endorsed Statement C and rejected D—again in agreement with essentialism—but the percentage decreased to 73%. Similarly, 76% of participants agreed with E and disagreed with F. The decrease may have been due to the complexity of the statements in C–F. The fact that participants had read in Story 12b that all cats, including Tibby, are robots controlled from Mars may have encouraged some of them to go along with D and F, because the second clause in each statement coincides with that information.

In the case of Story 12a, where a discovery is made about Tibby, 89% of participants accepted A and rejected B, supporting essentialist predictions. However, only 46% rejected both C and D, and only 47% rejected E and accepted F, which are the options Braisby et al. believed essentialists should take (see their Table 1). It is not obvious, however, that an essentialist would have to reject C. (Doesn't it count as a changed "belief about cats" that there are robots that look like cats?) Story 12a also begins with the statement that "you have a female pet cat named Tibby," which may have prompted the discrepant responses to E and F. (The percentages cited here are from Braisby et al., 1996, Experiment 2, which introduced procedural improvements over their Experiment 1.)

artifacts under the same circumstances (Ahn, 1998). Nor is there consistent evidence that explicitly describing the relation as causal has a greater impact on categorization than describing it as temporal or merely labeling it as a "dependency" (Sloman et al., 1998; but see Lassaline, 1996, for evidence that causal relations promote inductive generalization more powerfully than temporal ones). Finally, some studies have found that a missing cause has more impact than a missing effect (Ahn, 1998; Sloman et al., 1998), but others have not (Rehder & Hastie, 2001); in the latter study what mattered was the number of causal relations that an attribute enters into rather than the attribute's initial position in a causal network. Although causal relations clearly affect participants' categorization in all these studies—participants are less likely to classify an instance as a category member if it is missing an attribute that is part of a causal structure—it is unclear whether there is anything special about initial or internal causes.

These negative findings obviously need to be treated with caution, especially since these techniques are new ones. Different studies also use somewhat different methods. If we take the negative results seriously, however, they may provide a challenge to essentialism. Proponents might try to explain away these results on the grounds that artificial "natural" categories are not representative of natural natural kinds. Natural kinds just don't have causal structure in which, for example, functional properties cause molecular ones, so the way participants treat these items is irrelevant to the way they think about real kinds. But essentialists' hands are tied here by their commitment to psychological essentialism as a representation or pattern of beliefs. They cannot appeal to the way natural kinds really are to dismiss these experiments, particularly because some of the same theorists express doubts about scientific versions of essentialism (Gelman & Hirschfeld, 1999). If people believe that natural kinds have central causes that are important in producing many of the kinds' properties and in differentiating one kind from another, then why doesn't this show up as differences in performance in tasks that disrupt these assumptions? A better strategy for proponents of causal essentialism might be to maintain that the cover stories in these experiments are simply not convincing enough to engage assumptions about natural kinds in the first place. (Or, perhaps they are so convincing that they override people's everyday assumptions; B. Rehder, personal communication, August 3, 2001.) In addition, factors other than causal status can affect categorization, and it is possible that some of these overwhelm benefits due to initial or internal causes in some settings (Ahn & Kim, 2000). Of course, essentialists could simply accept the idea that what matters to people's ideas about natural kinds is the presence of causal forces and not whether the forces are internal to category members. But for reasons I take up later (see *The Intrinsic View as Beliefs About Natural Kinds* and *The Interaction View as Beliefs About Natural Kinds*), this may concede too much to competitors to essentialism.

Summary

The transformation studies quiz participants about whether a hypothetical change or discovery about an object prevents that object from being a member of a natural kind. Grade-school children and adults can perform these contrary-to-fact decisions, and they judge that certain changes to biological kinds (e.g., evisceration) do alter membership whereas other changes (e.g., external disguises) do not. Similarly, anyone who has had the usual dose of high school chemistry is likely to know, for example, that a hydrogen atom that has captured an extra proton is no longer hydrogen but something else (helium). They believe that atomic number differentiates the elements and is responsible for some of the elements' properties, but that the size, shape, texture, or color of a sample of the element does not. Those who think otherwise get very low grades.

Critics of the transformational studies point to limits on people's willingness to base their category decisions solely on central causal properties. Category membership may also depend on practical aspects of the natural kind—in the case of water, for example, where the substance is found and what people use it for. This may indicate some slippage between the scientific use of natural kind terms and our everyday use. For complex natural kinds (e.g., biological ones) that depend on many causal relations, people may believe that membership in the kind is graded rather than all or none. They may also think that the attributes most important for membership are ones that take part in many causal relationships rather than those that are the central cause (source of most effects) or those that are internal to the exemplars. These latter issues are not yet settled and warrant further investigation.

It also remains to be seen whether these limitations leave intact anything that could plausibly be called causal essentialism or psychological essentialism. This is a topic that I return to in the second part of this article. Nevertheless, the transformation studies make explicit what appears implicit in the induction experiments: People's knowledge of causal goings-on in natural kinds sustains inferences about what is possible for these kinds. In the case of the induction studies, causal relations help determine the range of properties that members of a kind can possess, given information about the properties of some of these members. In the transformation studies, causal relations help determine membership itself—the range of properties a member can possess and still be a member of that kind.

Natural Categories and Their Definitions

It's a truism among cognitive psychologists that people are unable to produce properties for natural categories that are both individually necessary and jointly sufficient for category membership. As we've seen, this is the reason proponents of causal essentialism reject "sortal essentialism" while maintaining that causal properties help differentiate natural kinds. The same idea can also be elevated to the level of a general principle (Fodor, 1981):

> Indeed, it seems to me to be among the most important findings of philosophical and psychological research over the last several hundred years (say, since Locke first made the reductionist program explicit) that attempts at conceptual analysis practically always fail. (p. 283)

The issue about necessary and sufficient properties raises questions about the types of properties that can legitimately play those roles. Any set of properties that are logically equivalent to *is a lion* is necessary and sufficient in one sense. For example, *being either a walnut or a lion* and *being a nonwalnut* are logically necessary and jointly sufficient for *being a lion,* but such properties are

surely not the sort that are relevant to psychological claims.[13] Traditionally, necessary and sufficient properties are supposed to be more primitive than the things they define, but it is far from obvious how to explicate this notion of primitiveness. Those who doubt the possibility of supplying necessary and sufficient properties believe that there is no reasonable way to spell out the primitiveness relation in such a way that the primitive properties successfully define the less primitive ones.

The published evidence on this issue is not completely one sided, however. Support for the truism comes from two studies in which participants explicitly listed properties for given categories, such as fish (Hampton, 1979), or listed properties for each of a set of subcategories (e.g., salmon, trout, and sardine) within the larger category (Rosch & Mervis, 1975). Judges then decided which of these properties apply to all and only members of that category—for example, whether there are any properties listed for fish that are true of all and only fish. The overall finding from these studies is that there are few listed properties common to all members, and those that are common also tend to apply to nonmembers. Thus, few if any properties are both necessary and sufficient. Hampton found, for example, that properties such as *is alive, lives in water,* and *is cold-blooded,* which his judges deemed true of all fish, are also true of nonfish, such as shrimp and tadpoles.

The results change, however, when the participants themselves, rather than the judges, do the labeling of properties as necessary and sufficient. McNamara and Sternberg (1983) asked participants to decide which of a set of properties were necessary for membership in specific natural kind and artifact categories (properties "exemplars of the word must have to be exemplars") and separately which properties or sets of properties were sufficient (properties that "guaranteed that some object or person was an exemplar of the given word"). On average, participants identified properties as both necessary and sufficient for 4.4 of 8 natural kind terms and 4.0 of 8 artifact terms that the experimenters presented. For example, participants identified the property *hardest substance known* as both necessary and sufficient for diamonds. Thus, McNamara and Sternberg (1983) concluded that "our investigations have led us to believe that the evidence against the definitional theories is less compelling than some have argued" (p. 470). These results are all the more surprising because McNamara and Sternberg's criteria for being necessary or sufficient are more stringent than Hampton's or Rosch and Mervis's: To qualify as necessary in Hampton's study, a property need only be true of all listed category members; in McNamara and Sternberg's, however, the property must be one that any member *must* have in order to be a member. It is difficult to tell whether the results of these studies differ because of differences in the stimulus categories, because of different standards adopted by judges versus participants, or because of other factors (see B. Tversky & Hemenway, 1984, and Murphy & Medin, 1985, for criticisms of property-listing methods).

Similar ambiguities surround evidence about the role of definitions in language understanding. In the 1970s and early 1980s, investigators used a number of methods to determine whether sentence comprehension and production are sensitive to the definitional complexity of individual words. The idea was that if (for example) *bachelor* can be defined as *unmarried man,* then sentences containing *bachelor* should be more complex than those containing *man.* Hence, if people have to translate the more complex, defined words into their simpler, undefined components to process such sentences, then sentences with *bachelor* should be more difficult to understand and to produce than ones with *man.* Experiments using sentence completion, sentence construction, word-relatedness judgments, and phoneme monitoring techniques turned up negative results on this score, using a variety of definitionally complex nouns and verbs (e.g., Fodor, Garrett, Walker, & Parkes, 1980; Kintsch, 1974, chap. 11). It is possible to take these results to suggest that there are no definitions—no singly necessary and jointly sufficient properties—for most natural language terms. However, a more cautious reading of the evidence is that effects of definitional complexity are rare in immediate language understanding and production. Positive effects are more common in tasks that require active inferences—for example, tasks where participants must determine the truth of a sentence on the basis of a picture (Clark, 1974; Just & Carpenter, 1971). We may need to recognize, then, that if necessary and sufficient predicates for a term exist, they may not be part of an entry in a mental lexicon whose definitions people consult whenever they encounter the term. But perhaps people store such information as facts about the term's referent, facts they can consult as needed in performing inferences and other tasks. (See McNamara & Miller, 1989, for a consideration of possibilities along these lines. See also Rips, 1995, for one way of drawing the distinction between representations of and about a category.)

It is reasonable to suppose that people have not only direct beliefs about properties of category members but also beliefs about the properties of such properties. They may think, for example, not only that fish are cold-blooded but also that cold-bloodedness is a necessary property of fish. They may also entertain these higher order beliefs in the absence of lower ones, believing that there are necessary and sufficient properties for natural categories, even though they are unsure of which properties fill this role (Malt, 1990; Medin & Ortony, 1989; Shipley, 1993). Such second-order beliefs about the origin and nature of natural kinds are frameworks for people's conceptions and are part of psychological essentialism, as we noted earlier. One source of evidence about these beliefs comes from a study by Malt (1990), who asked participants to consider objects that were in between two well-known categories—for example, a fish that "seems to you to be sort of halfway between" a sardine and an anchovy. Participants were to explain the category membership of the item by selecting one of the three choices in Option 13:

(13) a. It's probably one or the other, but I don't know which.
b. You can think of it as either one.
c. It can't really be either one, then.

For natural categories, such as sardines/anchovies, participants tended to select Option 13a, whereas for artifact categories (e.g., a vehicle that was between a car and a truck), participants chose Option 13b.

Not everyone subscribes to the existence of underlying objective criteria for natural kinds. Kalish (1995) found that participants are less likely to think there are facts that can settle membership in animal categories than in well-defined categories, such as odd

[13] Daniel Osherson has emphasized this point (personal communication, February 2001).

numbers. Nevertheless, in approximately 70% of trials participants said that disputes about membership in animal species could be settled by facts. People apparently believe, then, that there are criteria that decide membership in some natural categories, even when they themselves are unable to make the judgment. Likewise, even kindergartners apparently believe that a chimeric animal, midway between a chicken and a turkey, must be one or the other and not a hybrid (Keil, 1989, chap. 11).

Summary of Experimental Findings

Transformation studies make it clear that people think there are properties of objects that can affect their membership in natural kinds. Swap Leigh's genetic structure with that of a tiger, and she's no longer a lion but a tiger with a lion's appearance. But not all properties are relevant to membership in natural kinds. Painting stripes on Leigh does not change her status as a bona fide lion; she simply becomes a lion with a tiger's appearance.

How does this square with the difficulty in finding definitions for kind terms? If people know which properties are critical for lionhood and which are not, why can't they use such properties to define *lion*? Methodological differences among the studies may be partly responsible. Transformation studies sometimes ask participants whether certain named properties determine category membership, whereas experiments on necessary and sufficient properties have asked participants to produce their own property lists. It is certainly possible that people have more difficulty generating necessary and sufficient properties than recognizing such properties when they see them. Second, transformation studies often allow participants to be vague about the crucial properties in a way that is difficult or impossible in studies in which the participants must name the properties. It's one thing to know that something about a lion's insides makes it a lion and another to know exactly which thing is responsible. Similarly, knowing that cosmetic alterations don't change membership in natural categories doesn't entail knowing exactly which properties maintain membership. People's vagueness about these properties is one of the motivations for the view that people have only a placeholder for essential properties. Third, participants may think there is something wrong with listing a predicate like *having lion DNA* as a property of lions: *Having lion DNA* is circular in the sense of presupposing an understanding of lion. It is possible that *having lion DNA* does denote a necessary property of lions. Nevertheless, listing *lion DNA* may seem unhelpful in the same way listing *being a lion* is; these predicates provide no independent way of identifying the category in question. If participants see their task as providing properties that could aid someone in picking out lions, then *lion DNA* is useless if all one knows about it is that it is inside lions.

One could try to interpret the differences between the transformation and the definition studies by appealing to a distinction between causal and sortal essentialism, that is, between beliefs about kinds and beliefs about meaning. But the methodological variations I have just examined—in recognition versus production, in level of precision, and in epistemic or pragmatic demands—seem a more plausible explanation. The property-listing experiments that serve as evidence about lack of necessary and sufficient properties (e.g., Hampton, 1979; Rosch & Mervis, 1975) did not ask participants for definitions of terms per se. They asked instead for properties true of all and only category members. What is at stake in these experiments is participants' knowledge about lions, not their knowledge of lion definitions. This conclusion jibes with the earlier one (see the section *Questions About Transformation Studies*) that one should be careful about excluding evidence that appears on the surface to be about the meaning of kind terms rather than about the nature of natural kinds.

Of course, the evidence from definition and transformation experiments is not completely discordant. Malt's (1990) and Diesendruck and Gelman's (1999) experiments show that people sometimes believe there are facts of the matter about membership in kinds. Investigating an object in the proper way can in principle disclose properties about it that resolves its kind status. Thus, natural kinds possess a type of objectivity—natural kinds are kinds in nature—that serves as their hallmark and that differentiates them from more conventional kinds. McNamara and Sternberg's (1983) study suggests that people sometimes do identify properties of natural kinds that they think guarantee membership in the kind and that are necessary for continued existence in the kind. This tallies with evidence by Barton and Komatsu (1989) and others discussed in the context of the transformation studies. The explicitly modal character of these judgments is important, since it implies that people are basing their responses not merely on what is true of existing members of the category but on what is possible for them. To say that having a property is necessary for continued existence in a natural kind is not just a statement about the properties that all members of the kind happen to possess; it is a statement of the range of possibilities that are open to members. This is the same moral that emerged from the induction experiments: People's knowledge of natural kinds yields inferences about what is possible for these kinds. The central remaining question, then, is what psychological mechanisms could support such judgments?

What Explains Natural Categories' Modal Status?

The evidence I have reviewed shows that people are able to make judgments about natural kinds that extend beyond a tally of their members' properties. People can judge what would be true of members of these kinds, and what would then be true of members of related categories, under conditions that the kinds never in fact undergo. In the second part of this article, I evaluate some suggestions about what makes such judgments possible. The first section reviews cognitive theories of ideals and norms as ways of explaining beliefs about natural kinds' potential. The final section then turns to two traditional metaphysical views of the source of natural kinds' modal properties and asks whether a psychological version of one or the other could also serve as the basis of people's beliefs about kinds.

Do Norms or Ideals Underlie Beliefs About Kinds?

It seems possible that there are general psychological mechanisms, mechanisms that play a role in a variety of mental processes, that theorists might use to explain natural kinds' modal properties. Two immediate possibilities are those that Kahneman and Miller (1986; Kahneman, 1992; Kahneman & Varey, 1990) describe under the heading of *norms* and those that Barsalou (1985) calls *ideals*. Norms and ideals seem intrinsically normative and therefore may have the qualities we need to explain modal

beliefs about natural kinds. Moreover, these two proposals have some seemingly contrasting properties, which makes a comparison between them informative.

Ideals

Ideals are potential characteristics of objects that would best enable them to serve their category's goals. According to one of Barsalou's (1985) examples, foods with zero calories best fulfill the goal of the category of things to eat while on a diet, so having zero calories is an ideal for this category. Barsalou found that for explicitly goal-derived categories, such as diet foods, items that fell closer to the ideal (as determined by participant ratings) were also judged better examples of the category. It's sad, but true, that green tea (no sugar with that) is a better example of a diet food than are french fries. For other sorts of categories, including natural kinds and artifacts, closeness to ideals seemed to play some role in determining goodness of the example, but the effect was smaller than that of how close the item was to the category's central tendency. For instance, good examples of the category fruit are items like peaches and apples, which are highly similar to other fruit; however, good examples of fruit also include strawberries and bananas, which fall near the ideal of what people like to eat (but that are not especially similar to other fruit). In more recent research, Lynch, Coley, and Medin (2000) found that among tree experts the best examples of trees are those that are tallest and those that are least "weedy" (i.e., least messy). Central tendency played a lesser role in determining which trees are good examples.

Natural kinds sometimes serve human goals, and it is clear from Barsalou (1985), Lynch et al. (2000), and Malt (1994) that how well an object fulfills these goals can affect how typical it seems. Perhaps we can also regard as ideals not only properties that fulfill goals but also any extreme value on a dimension. In this sense, extreme height qualifies as an ideal for trees, even though extreme tree height may not be optimal for human purposes. Neither of these senses of ideal, however, is of much help in understanding the type of normativity we need in order to explain the induction and transformation studies. It seems probable that unfamiliar properties of apples or peaches are more likely to generalize to other fruits than are unfamiliar properties of bananas, despite the fact that they all fulfill human goals. Our knowledge of how properties of natural kinds are transmitted (and of how some properties depend on others) dictates that these patterns don't rely on how well their members subserve goals (unless, of course, the property is itself goal oriented). And, obviously, serving such goals is not likely to affect continued existence in the category. Many sorts of citruses are inedible, but they're fruits for all that. There may be exceptions in cases where humans have a role in creating the natural kinds, as in species of pets and crops. For these categories, breeders may have promoted certain properties or subspecies at the expense of others, given the breeders' goals. But such categories are more like artifacts than natural kinds and would produce different patterns of judgments in relevant induction and transformation studies.

Including among ideals those values that are extreme on some dimension (in addition to those that are optimal), does not help in explaining induction and transformation experiments. There is no reason to expect that members of natural kinds that are nearer extremes are more likely to share their other properties or that moderating these values would threaten their status as members. In isolated cases, extreme items may have greater inductive potential, as in the Itzaj example mentioned in the section *Category-Based Inductive Inference in Adults.* Some Itzaj believe that tall trees are more likely to transmit a novel disease than short ones. But this is because the causal path of transmission incidentally runs through the extreme instances—big trees come into physical contact with a larger number of other trees than do short ones—and not because of anything about extremity itself. In other Itzaj examples, culturally important species—for instance, jaguars—also carry greater inductive potential than species that are less prominent (Atran, 1998). However, there are plenty of other extreme values that presumably don't promote induction—for example, extremes of tree color or bark texture. Ideals are of great interest in the study of categories, especially categories of artifacts and goal-derived ensembles, and they probably affect performance in other cognitive spheres (Medin, Goldstone, & Markman, 1995). However, the normative qualities of ideals are not the qualities that are likely to explain the data reviewed earlier, at least for members of our own culture.

Norms and Mutability

According to Kahneman and Miller's (1986) approach, mention of an individual or a category produces a norm against which the original item is compared. The original individual or category will then seem surprising in retrospect to the extent that it is abnormal (i.e., differs from the constructed norm). The main theoretical proposal centers on how this norm is formed, a process that consists of two parts. First, the triggering individual or category recruits similar elements. A person gets these elements either by retrieving them from memory or by constructing them mentally, but in either case, the element is represented by a set of features or values of attributes. For example, seeing Leigh might bring to mind other lions you remember and perhaps other imagined lions that you've never actually experienced. The lions in this evoked set are each represented as a set of features, including, say, their height, length, color, furriness, and other properties. The second part of the norm-forming process puts together these evoked elements. According to Kahneman and Miller, people add the distribution of values for each property; therefore, if you remember several lions, then the distribution of lion heights will be added together, as will the distributions of colors, and so on. A value on each feature is considered normal if it is near maximum frequency and abnormal if it is near minimum frequency in these distributions.

It is clear that if the evoked set were confined to elements that a person has actually encountered and remembered, then the constructed norm would lack the properties required to explain the data reviewed earlier. The norm would simply reflect what is usual for observed category members. However, Kahneman and Miller (1986) embellished their idea of a norm with additional elements that bring it more in line with our requirements. First, they noted that in constructing a norm,

> the present model suggests that some features of the evoking stimulus are treated as immutable in that process: The recruitment of the evoked set tends to be restricted to elements that share these features.

> A plausible hypothesis is that the essential features that define the identity of the stimulus are most likely to be maintained as immutable. (p. 141)

Second, people can construct elements in the evoked set, as well as merely retrieving them, where the construction process also depends on which features of the triggering stimulus or category are mutable and which immutable.

So which features are immutable and which mutable? In some cases, this may depend on what a person wants to hold constant: "presuppositions are highly flexible, and the relative mutability of attributes can be controlled almost at will" (Kahneman & Miller, 1986, p. 143). In discussing the construction of elements, however, Kahneman and Miller hypothesized that the more mutable properties are those that are marked as exceptional, those that are less than ideal, those that are unreliable or less well known, those that are effects rather than causes, those that are at the center of attention, those that occur later in a series, and those that people can visualize. Two of these guiding principles, attentional focus and visualizability, don't seem powerful enough to guide thinking about natural kinds for the questions that concern us. (True, the more central properties of kinds may be the least easily visualizable ones, but mere difficulty in visualizing does not make them central.) I have also dealt with ideals in examining Barsalou's (1985) theory. Causal factors and reliability seem more important, in part because they are bound up with modal or counterfactual notions. Our understanding of what a cause is and what's reliable is inseparable from our beliefs about what events *would* change if others *were* to change. In particular, to determine what is possible for natural kinds, we do seem to take into account the causal network in which they are embedded (see the section *Direct Assessments of Causal Structure*). If this is correct, then the issue is how people deploy their beliefs about cause to project a natural kind's properties.[14]

Sloman et al. (1998) attempted to analyze mutability in terms of general dependency relations among the properties of a category. Central properties of a category member (e.g., a lion's genetic makeup) are those that many other properties depend on; peripheral properties (e.g., the shape of a lion's tail) are those that few other properties depend on. Thus, central properties are less mutable than peripheral ones. Sloman et al. did not analyze dependency further and believe that any sort of dependency relation affects mutability. This leaves it a little unclear, however, how we should understand the dependency concept. People probably believe that all and only animals with lion genes have lion tails. Hence, they believe in the "dependencies" *If an animal has lion genes, then the animal has a lion's tail* and *If an animal has a lion's tail, then it has lion genes.* Intuitively, however, only the first of these is the sort of dependency that is helpful in explaining natural kinds. Of course, participants' judgments in this study may well reflect the more crucial dependencies, but if so we need to know what criteria they use to distinguish the relevant dependencies from the irrelevant ones. This is another way of saying what Keil (1989) and others have noted earlier: There are too many dependencies—too many correlations among features, cue validities, category validities, and so on—to provide a principled explanation of natural kinds.

Two Frameworks for Natural Kinds

In view of the experimental findings, it is reasonable that beliefs about causal relations (and perhaps certain other dependencies) underwrite notions about what is possible for members of natural kinds. As Putnam (1990) put it, "To say that something is impossible is to say that nothing has the capacity to bring it about" (p. 74). This view, however, leaves many details to fill in. How exactly do causal beliefs shape our thinking about natural categories?

There are two main suggestions about natural kinds that are consistent with their modal standing and that may help in seeing what is at stake. According to one view, natural categories depend for their existence on certain internal properties of members that dictate category membership and many of their other (internal and external) properties. I call this the *intrinsic view.* The other tradition appears to be of more recent origin and focuses less on internal properties than on the causal role kinds play in relation to other things. I call this the *interaction view.* It is possible to trace the sources of these views, but the present goal is to sketch their main features, not to produce a historical account. These are not, of course, the only frameworks for thinking about natural kinds, but they're the ones most relevant for cognitive approaches. Most psychological theories of categories incorporate elements of both views as components of ordinary beliefs about kinds; both views present attractive features that are difficult to pass up. It is helpful to separate these views analytically, however, because running them together in an uncritical way produces confusions or contradictions. The contrast between them clarifies the claims of each.

In presenting these views, I start by describing them directly as theories about the structure of kinds, bypassing their potential as descriptions of people's beliefs. We can then see how well they fare in the role of psychological theories of natural categories.

The Intrinsic View

Imagine that there are a small number of properties intrinsic to certain objects that establish which natural kind the objects belong to. It is not easy to define precisely what is meant by *intrinsic,* but somehow the object's possessing the property does not depend on the existence of other objects. In particular, an object's intrinsic properties do not depend on us. Human interests and goals may be important in determining the way people classify artifacts, social kinds, and other categories, but except for the cases of human tampering mentioned earlier (see *Ideals* section), they don't affect natural kinds' intrinsic properties. People discover such properties, they don't invent or construct them, and the properties therefore have an objective status. Moreover, these properties are productive. Intrinsic properties of natural kinds are responsible for an unlimited number of other properties, so there is no end of information about them. These characteristics of intrinsic kinds coincide, then, with the first three properties in Table 1. In general, causal essentialism (the psychological theory) is the view that people *believe* natural kinds are intrinsic kinds; with this under-

[14] Once we have determined how this process works, however, we may be able to dispense with Kahneman and Miller's (1986) averaging over instances. The mechanism that produces counterfactual alternatives appears to do all the work in the cases being considered here.

standing, the Table 1 characteristics can do double duty as properties of intrinsic kinds and as properties of essentialist beliefs.

Essential properties are exactly those intrinsic properties that are crucial for natural-kind membership. Other intrinsic and extrinsic properties are accidental to membership. An essential property in this sense is not merely one that all and only members of a category happen to possess, but a property that an object *must* have to be a category member and, possibly, to exist at all. The set of essential properties is usually said to be "necessary and sufficient," but the relevant sense of these terms in this context is stronger than "every actual member has each essential property and every essential property is possessed by each actual member." Essential properties are properties that a member of the kind has across (possibly counterfactual) worlds or states or circumstances.[15] Moreover, essential properties make for distinctness between different kinds. If two objects belong to different natural kinds, then they must differ on at least one essential property. We therefore also have intrinsicness, uniqueness, distinctiveness, and identity from Table 1 as part of the intrinsic package.

The intrinsic view does not demand that arbitrary combinations of natural kinds are themselves natural kinds. There is no natural kind consisting of just daisies and aardvarks. However, some natural kinds may stand in subordinate–superordinate relations. In particular, according to the usual intrinsic view, if one object is a member of two different natural kinds, then one kind must be superordinate to the other. How many levels exist in the hierarchy of natural kinds is a question that has exercised philosophers and scientists from ancient to modern times (e.g., Atran, 1995, 1998; Lovejoy, 1936). The answer presumably depends on how many levels meet the criteria in Table 1. One possibility is that there is only a single level of intrinsic kinds, with higher and lower levels being more arbitrary sets. If so, this special level contributes the essential properties to its members. In Leigh's case, for example, the lion kind might be the one responsible for her essential properties; higher categories such as mammals and lower ones such as South Asian lions are then nonintrinsic kinds, perhaps imposed by people for classificatory convenience. The privileged level then decides the conditions of the object's existence. Take away the properties associated with being a lion and Leigh not only resigns from the lion category but ceases to exist entirely. On this strong version of the intrinsic story, "to be for a thing is to be a thing of a certain kind, to have a certain essence" (Loux, 1991, p. 7; see also Grene, 1963, p. 211, for a similar interpretation). Thus, intrinsic properties provide not only the essential properties of kinds but also the essential properties of individuals (Wiggins, 1980). On other weaker versions of the intrinsic story, there need be no special level, and several categories to which an object belongs provide criteria for sameness and persistence of the object, but only as a member of that kind. In this case, Leigh's membership in the lion category yields principles for determining whether something is the same lion as Leigh across circumstances (*identity of members* in Table 1), but not necessarily whether something is the same entity (i.e., *identity of individuals*).

On any version of the intrinsic view, though, it is the essence of a natural kind that is responsible for the modal characteristics of kind members. Essence determines what is possible for the member. Thus, if essence is some sort of internal causal force, as seems consistent with this view, then it is this causal essence that determines the limits on what a category member can do and be.

The Interaction View

The interaction view agrees with the intrinsic view in taking natural kinds to be objective and productive. The starting point for this view, however, is not the role of internal properties in the kinds but the role of the kinds in causal relations. Roughly speaking, natural kinds are the sorts of entities that causal laws relate. Chemical kinds, for example, are the sorts of things that participate in law-governed reactions with each other, and biological kinds the sorts of things that participate in law-governed biological relations involving reproduction, descent, and other matters (Fodor, 1974; Quine, 1969). The objectivity of natural kinds, on this view, then, is on a par with the objectivity of causal interactions that these laws describe. Similarly, to the extent that we can continue to discover new laws interrelating natural kinds, we can continue to discover new facts about the kinds themselves. The productivity of a natural kind's properties is on a par with the productivity of types of causal relations in which the kinds participate.

An object's membership in a natural kind depends on whether the object instantiates the laws for that kind, and preserving this relationship entails a harmony or equilibrium between laws and kinds. If certain objects or subcategories prove to be exceptions to a law involving the whole kind, then that may be a reason to suspect that the exceptions belong to some other kind. Alternatively, the "laws" that violate the integrity of a natural kind may be erroneous; the natural kind may participate in many other, better supported, laws that we might be unwilling to give up. This does not mean that kind membership is arbitrary, but it does mean that membership is more complex than the presence or absence of fixed internal properties. Table 2 summarizes this under the heading of *instantial membership,* which replaces the notion of the potency of essential properties (the method of establishing membership for intrinsic kinds in Table 1). For similar reasons, the existence of the entire category will depend on relational rather than purely intrinsic matters. Moreover, the interlocking of kinds and laws implies that natural kinds are projectible. Because causal laws and kinds are tailored to each other, natural kinds will support induction and counterfactual conditionals (Goodman, 1955). For example, if you learn that all known members of a natural kind have some property, *P,* then you can safely predict that a new member will also have *P,* provided that *P* is itself projectible. Under the same proviso, you can also suppose (counterfactually) that if something were a member of the same natural kind, then it too would have *P.*

Further differences between interactional and intrinsic kinds depend on assumptions about the (lack of) uniqueness and distinctiveness of a kind's causal relations. Nothing about interactional kinds demands that there be only one type of causal interaction (or only a very small number of types) that is crucial to the kind. In fact, I have just noted that the possibility of discovering new laws

[15] Proponents of intrinsic kinds hold that all essential properties are necessary and sufficient in this strong sense, but for reasons similar to those discussed earlier (see *Natural Categories and Their Definitions*), they may not hold the converse. There may be logically necessary and sufficient properties (e.g., *being a walnut or a lion* and *being a nonwalnut*) that are not essential, because they don't comport with other characteristics, such as potency and productivity (see Table 1). We can leave it as an open question for these proponents, however, how to separate true essential properties from others that are "merely" necessary and sufficient.

Table 2
Possible Characteristics of Interactional Kinds

Characteristic	Description
Instantial membership	Members of a natural kind are objects that instantiate the causal laws in which the kind participates.
Productivity	A kind's causal interactions are responsible for (a possibly unlimited number of) a member's properties.
Objectivity	Causal forces governing kinds exist in nature (do not depend on human convention).
Relationality	Existence of kinds depends on causal interactions with other objects or events.
Nonuniqueness	Natural kinds participate in many types of relevant causal relations. Individual members need not be part of each such relation.
Partial distinctiveness	Members of different natural kinds participate in different but possibly overlapping types of causal relationships.
Identity of members	Clusters of causal properties are responsible for tracing the same member of the kind across possible situations.
Identity of individuals?	Clusters of causal properties are responsible for tracing the same individual across possible situations.
Projectability	Natural kinds allow their properties to apply to new members and support counterfactual conditionals.

Note. As part of a psychological theory, the descriptions should be prefaced by "People believe that . . . ." Question mark indicates the characteristic considered optional.

is what is needed to explain natural kinds' productivity. Hence, the types of causal relations associated with an interactional kind are not unique in the way that essences are unique to intrinsic kinds. Moreover, although causal laws group objects in kinds, the causal interactions need not be common to all or only members of the kind, except as a limiting case (Wilson, 1999). A second sort of equilibrium prevails among a kind's causal properties: Subsets of the properties will causally support other subsets in overlapping and mutually reinforcing ways (Boyd, 1999; see also Keil, 1989, 1995, and Kornblith, 1993, who cite Boyd), and it is these stable subsets under which kinds coalesce. However, single members of the kind need not have all relevant causal properties. The causal properties are not discrete in the sense of Table 1—either all present or all absent in individual category members—and no single property will be prepotent in dominating the influence of all others. There is also nothing to prevent members of different kinds from having some of the same relevant causal properties. In this sense, then, causal properties only partially distinguish kinds (see Table 2). Of course, certain properties may turn out to be more central or less mutable than others, but this will depend on the particular configuration of causal forces that surround the kind.

Because of the interdependence of cause and kind, limits on the kind are limits on what the associated causes support. Questions about whether a kind can undergo certain changes and questions about the concomitants of such changes are questions about the causal interactions governing these transformations. If there are no essential properties, these issues are not decidable simply by inspecting whether intrinsic properties are preserved. Instead, one must defer to the same kind–cause harmony that determines membership in the first place.

The Intrinsic View as Beliefs About Natural Kinds

Let's suppose that people believe that natural kinds are intrinsic kinds. How does this stack up against the evidence reviewed thus far? At first glance, belief in intrinsic kinds comports well with results from the transformation experiments, which helped make the notion of psychological essentialism popular. Essential properties are those that can't be transformed away while still leaving an object's membership intact. So if children have a grip on which sorts of properties are essential for kinds and which are not, then they'll be able to appreciate that some transformations (i.e., those involving nonessential properties) preserve kind membership, whereas other transformations (i.e., those involving essential properties) do not. This follows from the potency and identity properties in Table 1. One qualification is that there is relatively little direct evidence that people can identify essential properties, but the intrinsic view is loosely consistent with this inability: Vague knowledge of these properties is sufficient for dealing with transformations but not for describing the properties precisely (as discussed in the *Summary* section for the first part of this article). Children and even adults may know the sorts of properties that are essential (e.g., they're inside an animal), but not precisely which properties these are. Perhaps a more serious qualification is the uncertainty surrounding evidence that intrinsic attributes count more toward membership than relational (functional) attributes in studies that have manipulated them (see *Direct Assessments of Causal Structure*). Absence of more direct evidence for essential properties is a weakness of the intrinsic view.

The intrinsic view explains some types of category-based induction. One of the theory's selling points is that because all members of a kind are alike with respect to essential properties, any property that is itself essential (or that depends heavily on the essential ones) will be similarly uniform across members. If such a property is true of one member, it must also hold of all others by a sort of "superinductive" inference (as Harper, 1989, calls it). There is no need to locate a convincingly large sample of members that possess the property before agreeing with the conclusion. A single instance may suffice (Nisbett et al., 1983). So far, so good for the intrinsic view. This advantage, however, should be balanced against the fact that essential properties must be supple-

mented with other mechanisms in order to explain typicality effects in inductive inference. It is unclear how essential properties could account for why people prefer generalizing from dogs to mammals over generalizing from opossums to mammals. Whatever essential properties mammals have are presumably shared equally by dogs and opossums. Essentialists might try to explain such findings by invoking similarity based on nonessential properties, as in earlier models of category-based induction. But my review of such findings suggests that people make use of deeper knowledge of causal relationships in making such judgments, not just surface similarity. Adding nonessential causal relations to the intrinsic view to handle these results is a concession to the rival interaction theory.

For counterfactual inductive inferences, the intrinsic view encounters more serious difficulties. Consider Argument 14, which seems reasonably strong and is the sort to which the gap model applies:

(14) Pekinese can leap over cars.

Dobermen can leap over cars.

To explain the inductive strength of Argument 14 via essential properties, it is necessary to assume that whatever essential properties Pekinese have are responsible for their leaping prowess. Suppose you think the premise is true (or are uncertain about its truth). It is plausible, then, that Dobermen have a similar set of essential properties that will also allow them to leap cars. But what if you think the premise is false (as you probably do)? In that case, to suppose the premise true for the sake of the argument, you either have to assume that Pekinese have acquired new essential properties or that something else, some other, nonessential causal mechanism, is responsible for their remarkable leaping skills in this counterfactual setting. However, the first alternative is out for intrinsic kinds: Essential properties are exactly those properties that *cannot* change across possible circumstances. Altering Pekinese's essential properties makes them ex-Pekinese, not Pekinese high jumpers. Changing essential properties of dogs, mammals, or animals has exactly the same consequence, as these properties are among the essential properties of Pekinese. Essential properties are useless, then, in explaining category-based induction with counterfactual premises. The same point can be put directly in terms of judgments about counterfactual conditionals. The statement *If Pekinese were able to leap over cars, then so could Dobermen* is presumably true, but not because Pekinese would have new essential properties in situations consistent with the antecedent.

The same problem holds for intrinsic kinds even if people have only "placeholders" for essential properties. If people believe that there is some unknown property P that is essential for Pekinese, then P must be true of Pekinese in at least all causally possible situations in which Pekinese exist—counterfactual as well as factual situations. For this reason, you can't assume for the sake of Argument 14 that Pekinese have adopted some new essential property P', which replaces P and enables them to jump higher. The very nature of essential properties implies that Pekinese do not exist without P, so it is incoherent to suppose the premise to be nonvacuously true under these new conditions.

This inability to explain category-based inferences through essential properties does not mean that the intrinsic view can't deal with such inferences at all. It is still possible to invoke other (nonessential) intrinsic properties or, for that matter, relational properties to explain the inductive strength of these arguments. But that means the intrinsic view ends up explaining modal phenomena in two different ways. On one hand, natural kinds' essential properties determine, for example, limits on the possible transformations that members can undergo and still count as members. On the other, some separate modal device is needed to explain possible effects that members have when counterfactual changes occur. Essential properties cannot be the sole source of the natural kinds' modal characteristics. This deficiency, together with the general lack of evidence that intrinsic properties are privileged in those studies that have manipulated causal relations explicitly (see *Direct Assessments of Causal Structure*), provides some grounds for skepticism about the intrinsic view as people's everyday theory about natural kinds.

The Interaction View as Beliefs About Natural Kinds

How does the interaction view compare? It is possible to take the interaction view as more general than the intrinsic view, because there is nothing to stop interactional kinds having some of their key causal properties essentially. So to make the comparison between the views sharper and more interesting, let's suppose that the interaction theory rules out essential properties. In this guise, the interaction view bears similarities to Rosch's (1978) theory: The view assumes that beliefs about natural kinds depend on clusterings of properties. Things with fins tend to have gills, things with feathers tend to have wings, and so on. According to the interactional view, however, causal relations replace mere statistical co-occurrences. Transformation experiments might seem at first to be weak spots for this view, as a few of these studies (e.g., Barton & Komatsu, 1989; Gelman & Wellman, 1991) suggest that people think a change to certain properties entails a change in kind membership. Aren't the changed properties essential ones in these circumstances?

However, properties that change kind are not necessarily essential properties. If the interaction view is correct, membership in a natural kind may depend on many interacting properties. If enough of these properties are disrupted, for example, by eviscerating an organism, then the organism is no longer a category member. But no single property (or small subset of properties) need be essential in this situation. The fact that interactional kinds have no essential properties, then, does not mean that there is no way in which members of these kinds can cease to be members. Objects are members of interactional kinds in virtue of instantiating causal laws (see Table 2). So if a member begins violating enough of these laws (whereas other members continue to instantiate them), then we should consider the object no longer a member of the kind in question. Of course, a proponent of intrinsic kinds could *stipulate* that any property or group of properties that cause a change in natural category is thereby an essential property, but this would sacrifice other characteristics of Table 1 (e.g., uniqueness and distinctiveness) that are part of the intrinsic view's appeal. This stipulation might be a route to a compromise between intrinsic and interaction views (and it may even be what is behind psychological essentialists' rejection of "sortal essentialism"), but it makes the idea of an essential property less clear cut, isolating it from its usual matrix.

The interaction view can also handle category-based induction. Individual members of interactional kinds have overlapping causal properties; hence new properties that are bound up with the old ones will tend to carry over to other members of the same kind. Overlap in these properties will not guarantee that a new biological property that Leigh has—say, having Enzyme E—will hold for all lions. But, in fact, we are often not certain about such matters. Whether other lions have the enzyme depends on how the property is hooked up to other properties of the kind and on the connections among these mediating properties, as discussed earlier (see sections on category-based induction). Thus, the interaction view coincides with intuition on this matter.

The interaction view accounts for inferences based on counterfactual information in a uniform way. To accommodate the counterfactual, people have to suppose that the properties of a kind (or of a member of the kind) differ from what they are in the current state of affairs, making adjustments in causally dependent properties. The strength of the inference will then depend on what these adjustments lead us to think will be true in the adjusted state. One advantage of interactional kinds, then, is parsimony: Why posit two modal mechanisms when only one appears necessary? There may, of course, be considerations that would speak in favor of two separate sources for modal phenomena. Perhaps theories of reference could provide reasons for a dual theory of this sort. However, if the only truly essential properties such theories guarantee are ones like bearing-the-same-kind-relation-to-local-samples, then they don't motivate the full-blown essentialism outlined in Table 1.

The interaction view also gets for free the general lack of evidence for definitions, because there are no essential properties to supply the definitions. In those few studies reporting that people do provide definitions for natural kinds, the interaction view can interpret these as cases in which the participants report important, perhaps even universal, properties but not essential ones. This seems consistent with the actual examples that these studies cite. For example, being the hardest substance known is true of diamonds, but it is not essential because it leaves open the possibility of harder unknown substances. A potential difficulty for interactional kinds, however, is Malt's (1990) finding that people think that an object in between two natural kinds is "probably one or the other." A crisp division between kinds would support the intrinsic view's distinctiveness over the interaction view's partial distinctiveness (see Tables 1 and 2). The strength of this objection, of course, depends on the reliability of the effect (Kalish, 1995) and on issues of wording—for example, on how hard participants leaned on "probably" in "probably one or the other." It is consistent with the interaction view that the causal forces responsible for kinds introduce a fairly clear separation between them, making in-between cases "improbable."

A second potential disadvantage of interactional kinds is that the tight connection between kinds and laws may put natural kinds out of the reach of children and adults who don't know the relevant science. Certainly, young children don't know what types of chemical laws govern kinds like water, and neither did adults before modern times. It is therefore tempting to say that these people cannot have interactional kinds, even though they appreciate water as a category distinct from other liquids. Maybe early kinds are intrinsic kinds, whose status changes to an interactional kind if a person happens to acquire the relevant scientific principles. Children might start out with a simplified causal schema that specifies a unique, distinctive organizing cause per kind and later graduate to a more differentiated set of laws that provide the basis for an interactional, theory-based kind. A number of authors (e.g., Keil, 1989; Kornblith, 1993) have pointed out that an innate psychological bias toward expecting intrinsic kinds may help children acquire later knowledge of interactive ones.

There is no strong reason, though, for proponents of interactional kinds to concede that early kinds are intrinsic. Clearly, beliefs about kinds change as we add knowledge about laws that govern the kinds. However, it is uncertain whether the start state for learning consists of belief in a single internal cause or belief in multiple—possibly fragmentary, possibly false—interactive ones. It seems at least as plausible to suppose that children begin by assuming that what makes Leigh a lion is a set of causes (e.g., having lion cubs, having a lion mom, dominating other animals) as a single intrinsic cause. Skepticism about interactional kinds seems to arise from doubts about children's knowledge of causal principles, but the lawlike character of natural kinds is also part of the intrinsic story. The intrinsic view can soften its position by positing that people have beliefs in such laws without knowing the laws' descriptive content, but the same move is open to proponents of interactional kinds. Similarly, if an innate bias toward intrinsic kinds fosters later knowledge of science, a bias toward interactional kinds should do at least as well.

The most serious complaint about interactional kinds is that they are too unconstrained, especially if they amount to no more than the bland notion that natural categories have causes.[16] Intrinsic kinds come with a commitment to unique central causes that produce an organism's surface characteristics. Thus, intrinsic kinds explain the impression that their members distribute statistically around a common type, and they motivate a search for the type's source. By contrast, interactional kinds seem less disciplined, as there are multiple forces that shape the kinds. Still, the interactional view can hold that natural kinds result not just from any combination of forces but from special conditions that yield instantial membership, productivity, identity, partial distinctiveness, and the other characteristics in Table 2. (Evolutionary views of biological kinds provide an expert version of this type of thinking.) One might also question whether positing a single root cause is the best or most intuitive explanation for distributional characteristics. Intrinsic kinds must posit interfering forces to explain why category members aren't cookie-cutter versions of each other. Interactional causes seem an equally reasonable source of variability. Certainly, the ecological reasoning that appears in cross-cultural studies of category-based induction (e.g., Lopez et al., 1997; Coley et al., 1999; Medin et al., 1997; Proffitt et al., 2000) is more at home in an interactional than in an intrinsic context. Likewise, although belief in a central cause can drive inquiry in everyday life and in science, so could belief in cooperating causes. If we discovered that there were no essential causes

[16] This point is due to Douglas Medin (personal communication, October 11, 2000).

for plant and animal species, would that squelch our curiosity about their origins and properties?[17]

Summary and Concluding Comments

People have knowledge of what is possible for members of natural kinds, not just of what is presently true of them. It is unclear to what extent people can volunteer properties that objects must have in order to be members of these kinds, but they make consistent judgments about the membership of hypothetical objects that have gained or lost properties, as well as similar judgments about what these hypothetical changes imply for other kinds of objects. How is this knowledge of the possible possible?

The second part of this article examined whether causal information could provide the basis for natural kinds' modal properties. Knowledge about causes may be helpful in this regard because such principles apply not just to natural kinds as they currently are but also to kinds in other potential situations. To be more specific about the kind–cause relation, however, investigators need to know how people think causal forces shape natural kinds. One possibility along these lines is that kinds depend on a single, intrinsic cause that is responsible for an object's membership, typical properties, identity, and distinctness from other kinds. This is the intrinsic view that Table 1 summarizes, and it captures current ideas about causal (psychological) essentialism. An alternative possibility, outlined in Table 2, sees natural kinds as constellations of causal forces. Both conceptions take natural kinds to be objective groupings "in nature" with a potentially unlimited number of properties to be discovered, but interactional kinds depend on a set of meshing causes rather than a unique internal essence.

There may be room for both kinds of kinds. For example, adults might believe that chemical kinds are intrinsic, but biological kinds interactional.[18] Intrinsic kinds seem better equipped to handle judgments that membership in natural kinds is all or none. Interactional kinds have the advantage of explaining the inductive and transformation data in a unified way. Intrinsic kinds force us to distinguish between essential properties responsible for change in membership and nonessential, but modal, properties responsible for other counterfactual changes; interactional kinds can do both jobs with the same causal forces.

[17] It is also good to notice that both intrinsic and interactional kinds pose problems of circularity. There is a potential explanatory circle running between a natural kind and its essence (Q: What's a lion? A: Whatever is caused by lion essence. Q: What's lion essence? A: Whatever causes lions.), and a similar circle runs between natural kinds and the causal laws in which it participates. People might be content with such circular beliefs for awhile, but very tight circles are likely to seem feeble eventually.

[18] See Sober (1980) for a metaphysical view along these lines.

References

Abbott, B. (1997). A note on the nature of "water." *Mind, 106,* 311–319.

Ahn, W. (1998). Why are different features central for natural kinds and artifacts? The role of causal status in determining feature centrality. *Cognition, 69,* 135–178.

Ahn, W., & Kim, N. (2000). The causal status effect in categorization: An overview. *Psychology of Learning and Motivation, 40,* 23–65.

Atran, S. (1995). Classifying nature across cultures. In E. E. Smith & D. N. Osherson (Eds.), *Thinking* (2nd ed., pp. 131–174). Cambridge, MA: MIT Press.

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences, 21,* 547–609.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 629–654.

Barton, M. E., & Komatsu, L. K. (1989). Defining features of natural kinds and artifacts. *Journal of Psycholinguistic Research, 18,* 433–447.

Blok, S., Newman, G., Behr, J., & Rips, L. J. (2001). Inferences about personal identity. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society,* 80–85.

Boyd, R. (1999). Homeostasis, species, and higher taxa. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 141–185). Cambridge, MA: MIT Press.

Braisby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition, 59,* 247–274.

Brandom, R. (1988). Inference, expression, and induction. *Philosophical Studies, 54,* 257–285.

Brandom, R. (1994). *Making it explicit.* Cambridge, MA: Harvard University Press.

Burstein, M. H., Collins, A., & Baker, M. (1991). Plausible generalization: Extending a model of human plausible reasoning. *Journal of the Learning Sciences, 1,* 319–359.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Carey, S. (1995). On the origins of causal understanding. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition* (pp. 268–302). Oxford, England: Oxford University Press.

Carnap, R. (1950). *Logical foundation of probability.* Chicago: University of Chicago Press.

Clark, H. H. (1974). *Semantics and comprehension.* The Hague, the Netherlands: Mouton.

Coley, J. D., Medin, D. L., & Atran, S. (1998). Does rank have its privilege? Inductive inferences within folkbiological taxonomies. *Cognition, 64,* 73–112.

Coley, J. D., Medin, D. L., Proffitt, J. B., Lynch, E., & Atran, S. (1999). Inductive reasoning in folkbiological thought. In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 205–232). Cambridge, MA: MIT Press.

Collins, A., & Michalski, R. (1989). The logic of plausible reasoning: A core theory. *Cognitive Science, 13,* 1–49.

Diesendruck, G., & Gelman, S. A. (1999). Domain differences in absolute judgments of category membership. *Psychonomic Bulletin and Review, 6,* 338–346.

Dupré, J. (1993). *The disorder of things: Metaphysical foundations of the disunity of science.* Cambridge, MA: Harvard University Press.

Ellis, B. (1996). Natural kinds and natural kind reasoning. In P. J. Riggs (Ed.), *Natural kinds, laws of nature, and scientific methodology* (pp. 11–28). Dordrecht, the Netherlands: Kluwer Academic.

Fodor, J. A. (1974). Special sciences. *Synthese, 28,* 77–115.

Fodor, J. A. (1981). The present status of the innateness controversy. In *Representations* (pp. 257–316). Cambridge, MA: MIT Press.

Fodor, J. A., Garrett, M. F., Walker, E. C. T., & Parkes, C. H. (1980). Against definition. *Cognition, 8,* 263–367.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology, 20,* 65–95.

Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird. *Developmental Psychology, 26,* 796–804.

Gelman, S. A., & Hirschfeld, L. A. (1999). How biological is essentialism? In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 403–446). Cambridge, MA: MIT Press.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23,* 183–209.

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 essences: Early understanding of the nonobvious. *Cognition, 38,* 213–244.

Ginsberg, M. L. (1987). Introduction. In M. L. Ginsberg (Ed.), *Readings in nonmonotonic reasoning* (pp. 1–23). Los Altos, CA: Morgan Kaufmann.

Goodman, N. (1955). *Fact, fiction, and forecast.* Cambridge, MA: Harvard University Press.

Grene, M. (1963). *A portrait of Aristotle.* London: Faber and Faber.

Gutheil, G., & Gelman, S. A. (1997). Children's use of sample size and diversity information with basic-level categories. *Journal of Experimental Child Psychology, 64,* 159–174.

Hall, D. G. (1998). Continuity and the persistence of objects: When the whole is greater than the sum of the parts. *Cognitive Psychology, 37,* 28–59.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18,* 441–461.

Hampton, J. A. (1995a). Similarity-based categorization: The development of prototype theory. *Psychologica Belgica, 35,* 103–125.

Hampton, J. A. (1995b). Testing the prototype theory of concepts. *Journal of Memory and Language, 34,* 686–708.

Harper, W. (1989). Consilience and natural kind reasoning. In J. R. Brown & J. Mittelstrass (Eds.), *An intimate relation* (pp. 115–152). Dordrecht, the Netherlands: Kluwer.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford, England: Oxford University Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7,* 569–592.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 411–422.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

Hobbes, T. (1839–1845). De corpore. In W. Molesworth (Ed.), *The English works of Thomas Hobbes* (Vol. 1). London: John Bohn.

Inagaki, K., & Hatano, G. (1993). Young children's understanding of the mind-body distinction. *Child Development, 64,* 1534–1549.

Johnson, S. C., & Carey, S. (1998). Knowledge enrichment and conceptual change in folk biology: Evidence from people with Williams syndrome. *Cognitive Psychology, 37,* 156–200.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior, 10,* 244–253.

Kahneman, D. (1992). Reference points, anchors, norms, and mixed feelings. *Organizational Behavior & Human Decision Processes, 51,* 296–312.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93,* 136–153.

Kahneman, D., & Varey, C. A. (1990). Propensities and counterfactuals: The loser that almost won. *Journal of Personality and Social Psychology, 59,* 1101–1110.

Kalish, C. W. (1995). Essentialism and graded membership in animal and artifact categories. *Memory & Cognition, 23,* 335–349.

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

Keil, F. C. (1995). The growth of causal understanding of natural kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition* (pp. 234–262). Oxford, England: Oxford University Press.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition, 65,* 103–135.

Kintsch, W. (1974). *The representation of meaning in memory.* Hillsdale, NJ: Erlbaum.

Kornblith, H. (1993). *Inductive inference and its natural ground.* Cambridge, MA: MIT Press.

Krifka, M., Pelletier, F. J., Carlson, G. N., ter Meulen, A., Chierchia, G., & Link, G. (1995). Genericity: An introduction. In G. N. Carlson & F. J. Pelletier (Eds.), *The generic book* (pp. 1–124). Chicago: University of Chicago Press.

Kripke, S. A. (1972). *Naming and necessity.* Cambridge, MA: Harvard University Press.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 754–770.

Levi, I. (1996). *For the sake of the argument.* Cambridge, England: Cambridge University Press.

Liittschwager, J. C. (1994). Children's reasoning about identity across transformation. *Dissertation Abstracts International, 55*(10), 4623B. (University Microfilms No. AAC95-08399)

Lopez, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology, 32,* 251–295.

Loux, M. J. (1991). *Primary ousia: An essay on Aristotle's Metaphysics Z and H.* Ithaca, NY: Cornell University Press.

Lovejoy, A. O. (1936). *The great chain of being: A study of the history of an idea.* Cambridge, MA: Harvard University Press.

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

Malt, B. C. (1990). Features and beliefs in the mental representations of categories. *Journal of Memory and Language, 29,* 289–315.

Malt, B. C. (1994). Water is not H_2O. *Cognitive Psychology, 27,* 41–70.

Mandler, J. M., & McDonough, L. (1998). Studies in inductive inference in infancy. *Cognitive Psychology, 37,* 60–97.

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

McCloskey, M., & Glucksberg, S. (1978). Natural categories: Well-defined or fuzzy sets? *Memory & Cognition, 6,* 462–472.

McDonald, J., Samuels, M., & Rispoli, J. (1996). A hypothesis-assessment model of categorical argument strength. *Cognition, 59,* 199–217.

McNamara, T. P., & Miller, D. L. (1989). Attributes of theories of meaning. *Psychological Bulletin, 106,* 355–376.

McNamara, T. P., & Sternberg, R. J. (1983). Mental models of word meaning. *Journal of Verbal Learning and Verbal Behavior, 22,* 449–474.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 1469–1481.

Medin, D. L., Goldstone, R. L., & Markman, A. B. (1995). Comparison and choice: Relations between similarity processes and decision processes. *Psychonomic Bulletin & Review, 2,* 1–19.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts. *Cognitive Psychology, 32,* 49–96.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge, England: Cambridge University Press.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90,* 339–363.

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

Osherson, D. N., Smith, E. E., Myers, T. S., Shafir, E., & Stob, M. (1994). Extrapolating human probability judgment. *Theory and Decision, 36,* 103–129.

Osherson, D. N., Smith, E. E., & Shafir, E. B. (1986). Some origins of belief. *Cognition, 24,* 197–224.

Osherson, D. N., Smith, E. E., Shafir, E., Gualtierotti, A., & Biolsi, K. (1995). A source of Bayesian priors. *Cognitive Science, 19,* 377–405.

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

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 811–828.

Putnam, H. (1975). The meaning of "meaning." In K. Gunderson (Ed.), *Language, mind, and knowledge* (pp. 131–193). Minneapolis: University of Minnesota Press.

Putnam, H. (1990). Is water necessarily H_2O? In J. Conant (Ed.), *Realism with a human face* (pp. 54–79). Cambridge, MA: Harvard University Press.

Quine, W. V. (1969). Natural kinds. In *Ontological relativity and other essays* (pp. 114–138). New York: Columbia University Press.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 382–407.

Rehder, B., & Hastie, R (2001). The essence of categories: The effects of underlying causal mechanisms on induction, categorization, and similarity. *Journal of Experimental Psychology: General, 130,* 323–360.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14,* 665–681.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge, England: Cambridge University Press.

Rips, L. J. (1995). The current status of research on concept combination. *Mind and Language, 10,* 72–104.

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

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

Rosengren, K. S., Gelman, S. A., Kalish, C. W., & McCormick, M. (1991). As time goes by: Children's early understanding of growth in animals. *Child Development, 62,* 1302–1320.

Ross, B. H., & Murphy, G. L. (1999). Food for thought. *Cognitive Psychology, 38,* 495–553.

Shipley, E. F. (1993). Categories, hierarchies, and induction. *Psychology of Learning and Motivation, 30,* 265–301.

Simons, D. J., & Keil, F. C. (1995). An abstract to concrete shift in the development of biological thought: The *insides* story. *Cognition, 56,* 129–163.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology, 25,* 231–280.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition, 52,* 1–21.

Sloman, S. A. (1997). Explanatory coherence and the induction of properties. *Thinking & Reasoning, 3,* 81–110.

Sloman, S. A., & Ahn, W.-K. (1999). Feature centrality: naming versus imagining. *Memory & Cognition, 27,* 526–537.

Sloman, S. A., Love, B. C., & Ahn, W.-K. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22,* 189–228.

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science, 12,* 485–527.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition, 49,* 67–96.

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition, 22,* 377–386.

Sober, E. (1980). Evolution, population thinking, and essentialism. *Philosophy of Science, 47,* 350–383.

Solomon, G. E. A., Johnson, S. C., Zaitchik, D., & Carey, S. (1996). Like father, like son: Young children's understanding of how and why offspring resemble their parents. *Child Development, 67,* 151–171.

Springer, K. (1996). Young children's understanding of a biological basis for parent-offspring relations. *Child Development, 67,* 2841–2856.

Springer, K., & Keil, F. C. (1989). On the development of biologically specific beliefs: The case of inheritance. *Child Development, 60,* 637–648.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–112). Oxford, England: Blackwell.

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition, 74,* 149–175.

Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. *Proceedings of the 23rd annual Conference of the Cognitive Science Society,* 1036–1041.

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

Tversky, A., & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General, 113,* 169–193.

Waxman, S. R., Lynch, E. B., Casey, K. L., & Baer, L. (1997). Setters and samoyeds: The emergence of subordinate level categories as a basis for inductive inference in preschool-age children. *Developmental Psychology, 33,* 1074–1090.

Wiggins, D. (1980). *Sameness and substance.* Cambridge, MA: Harvard University Press.

Wilson, R. A. (1999). Realism, essence, and kind: Resuscitating species essentialism? In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 187–207). Cambridge, MA: MIT Press.

Received October 10, 2000
Revision received April 2, 2001
Accepted April 16, 2001 ■