Psychonomic Bulletin & Review
2000, 7 (4), 569-592

Properties of inductive reasoning

EVAN HEIT
University of Warwick, Coventry, England

This paper reviews the main psychological phenomena of inductive reasoning, covering 25 years of experimental and model-based research, in particular addressing four questions. First, what makes a case or event generalizable to other cases? Second, what makes a set of cases generalizable? Third, what makes a property or predicate projectable? Fourth, how do psychological models of induction address these results? The key results in inductive reasoning are outlined, and several recent models, including a new Bayesian account, are evaluated with respect to these results. In addition, future directions for experimental and model-based work are proposed.

Imagine that during an evening while you are out at the theater, your home is broken into and several personal items are stolen. This sudden event, in addition to having practical and possibly emotional consequences, is going to lead to changes in your beliefs and predictions about the future. Whereas you may have previously thought that your home was secure, you may now believe, on the basis of this one event, that it is rather likely that your home will be burgled again.

In the terms of inductive reasoning, you may well see similarities between one case—your home on this particular evening—and future cases—that is, your home on other, future evenings—leading you to project a predicate—being burgled—from the one case to the others. Of course, carrying out this sort of inductive reasoning would be more complicated, because there are the many past cases of evenings on which your home has not been burgled, and these cases too seem to have implications for the future. In addition, other information may be useful, such as whether or not nearby homes have been burgled recently. It seems that due to the similarity in location, knowing the history of other homes would help you to predict the safety of your own home.

This paper addresses how people project information from known cases to the unknown. The aim is to integrate the findings from a large number of psychological studies conducted over the past 25 years, on adults as well as children. From a tradition starting with Rips (1975), psychological experiments on inductive reasoning have typically addressed how people make inferences about predicates or properties of things such as animals—for example, about whether a dog is susceptable to a particular kind of disease—rather than idiosyncratic events such as home burglaries. One reason for the extensive study of reasoning using animal categories rather than individual personal events is that we have a rich and well-documented categorical structure for representing animals and other living things.

It is possible to think of many cognitive activities as containing an element of inductive reasoning, using the known to predict the unknown; such activities range from problem solving to social interaction to motor control. However, this paper will focus on a narrower range of phenomena, concerning how people evaluate inductive arguments such as the following example:

Goldfish thrive in sunlight

Tunas thrive in sunlight.

The information above the line is taken as a premise that is assumed to be true; the task is to evaluate the likelihood or strength of the conclusion, below the line. There are several possible variants of this task. For example, the premise and conclusion could be presented as sentences or in pictures. There could be more than one premise. In addition, information in the premises could be provided for a category, such as all goldfish, or for an individual, such as one particular fish. Likewise, the conclusion could refer to a category or a specific individual. Finally, there are several ways to collect judgments about the conclusion; one could, for example, require responses on a scale of probability or inductive strength, or forced-choice judgments, in which subjects must choose between different conclusions. Indeed, some studies could be described as collecting behavioral judgments rather than asking questions. For example, in some infant studies, induction is measured in terms of what action the child performs with a particular toy. Generally speaking, not all the results reported in this paper have been documented for all the different task variants, because researchers have typically assumed that the different variants address the same underlying processing. However, when system-

I am grateful for comments from Beth Proffitt, John Coley, Ulrike Hahn, Philip Johnson-Laird, Jean Mandler, Dan Osherson, and Elizabeth Shipley. Much of this material was first presented as a series of lectures at the Sixth International Summer School in Cognitive Science, New Bulgarian University, Sofia, 1999. This research has benefited from funding by the Biotechnology and Biological Sciences Research Council, and the Economic and Social Research Council, of the United Kingdom. Please address correspondence to E. Heit, Department of Psychology, University of Warwick, Coventry CV4 7AL, England (email: e.heit@warwick.ac.uk).

Table 1
Touchstone Results in Inductive Reasoning

Inferences From Single Cases

1. Similarity between premise and conclusion categories promotes induction.
2. Typicality of the premise category promotes induction. (No corresponding findings for the conclusion category.)
3. Homogeneity of the conclusion category promotes induction. (No corresponding findings for the premise category.)

Inferences From Multiple Cases

4. Greater number of observations, or premises, promotes induction (although the evidence is weak for children).
5. Greater diversity of observations, or premises, promotes induction (although the evidence is mixed for children, and too much diversity may not help even for adults).

Influence of Properties

6. There is widespread evidence that people draw inferences differently depending on the property being projected (found in adults and children).
7. Some properties are idiosyncratic or transient, with a narrow scope for inferences, whereas other properties are more broadly projected.
8. The assessment of similarity between categories in an argument depends on the property being projected.

atic differences between different task versions have been reported, these will be highlighted.

This paper is intended to answer a number of questions about inductive reasoning, using the current findings from psychological research. The first three questions are factual and empirical, concerning how people respond to various kinds of inductive arguments. First, what makes a case generalizable? That is, when does an observation that something has a certain property promote the inference that something else has that property? Second, what makes a set of cases generalizable? The evidence shows that simply putting together a list of the most convincing, or induction-promoting, cases does not necessarily lead to the strongest possible ensemble of cases. The interesting result is that sometimes a set of individually weak cases can make a strong case together. Third, what makes a property projectable? That is, when we observe an object with various properties, which properties of the object are more likely to be projected to another case or inferred than others? Many psychological studies of inductive reasoning have addressed more than one of these questions. Therefore, different facets of the results of these studies will be described at different points in this paper.

The final question to be addressed is as follows: What are the psychological models of inductive reasoning? In the fourth main section of this paper, formal models of inductive reasoning will be discussed. Rather than present all of these accounts in detail, these accounts will be described just in terms of how they address the results covered by the first three questions. In Table 1, the touchstone results from psychological experiments on inductive reasoning will be listed, and the models of induction will be assessed against this list.

However, before proceeding with this review of psychological work on inductive reasoning, it is worth acknowledging that the study of induction has a longer history in other fields such as philosophy. Perhaps the best-known analysis from philosophy is Hume's (1748/1988) argument against the logical justification of induction. Hume argued that, unlike deductive inference, there is no basis for establishing the validity of a method for drawing inductive inferences. Although psychological work on induction has not directly addressed this traditional problem of induction, psychological research does paint a somewhat more optimistic picture, emphasizing how inductive reasoning is widespread in human thought and how people perform this reasoning very systematically. Psychological research has uncovered a rich and interesting set of phenomena that reveal much about cognitive processes. Furthermore, although they fall short of a complete logical justification for induction, some psychological accounts have addressed whether people's patterns of inductive reasoning do meet basic cognitive goals and to what extent people are subject to fallacies or internal contradictions. As the psychological phenomena are reviewed in this paper, when there have been related philosophical analyses these will be presented as well.

What makes a good case?

The first issue to be addressed is why do we more readily draw inferences from some cases than others? For example, hearing about a burglary 2 miles away normally would have more effect than a burglary 100 miles away, on inferences about the security of one's own home. In fact, the notion of proximity is central to understanding induction, because similarity between cases has been found to be one of the main determinants of inductive strength. Actually, in this section, two questions will be covered. First, what is it about a premise category that promotes inferences to a conclusion category? Second, what makes a conclusion category itself seem like a good

target for inferences? All of the results in this section will refer to situations where there is a single premise provided, because there is a well-established set of central phenomena for single-premise arguments. In the next section of this paper, inferences using multiple premises will be considered.

Initial adult studies. The seminal study of inductive reasoning was that of Rips (1975). This work looked at how adults project properties of one category of animals to another. Subjects were told to assume that on a small island, it had been discovered that all members of a particular species (of birds or mammals) had a new type of contagious disease. Then the subjects judged for various other species what proportion would also have the disease. For example, if all rabbits had this disease, what proportion of dogs would have it? Rips used a variety of animal categories in the premise and conclusion roles, with the categories having a known similarity structure derived using multidimensional scaling techniques. It was found that two factors consistently promoted inferences from a premise category to a conclusion category.

First, similarity between premises and conclusions promoted strong inferences. For example, subjects made stronger inferences from rabbits to dogs than from rabbits to bears. Second, the typicality of the premise, with respect to its superordinate category, was critical in promoting inferences. (Typicality of *rabbit*, for example, would be measured in terms of its distance from the representation of its superordinate, *mammal*, in a multidimensional scaling solution.) The result was that more typical premise categories led to stronger inferences than did atypical premise categories. For example, with the bird stimuli, having *bluejay* as a premise category led to stronger inferences overall than did having *goose* as a premise category. Using multiple regression analyses, Rips (1975) found distinct contributions of premise–conclusion similarity and premise typicality. Interestingly, there was no evidence for a role of conclusion typicality. For example, all other things being equal, people would be as willing to draw a conclusion about a bluejay or about a goose, despite the difference in typicality of these two categories. It is important to keep these three findings in mind, because they recur in many subsequent studies of inductive reasoning—namely, that premise–conclusion similarity and premise typicality promote induction, but that typicality of the conclusion category does not seem to affect inductive strength.

Chronologically speaking, the next major study in this paradigm was done by Nisbett, Krantz, Jepson, and Kunda (1983), who also asked subjects to draw inferences about items (animals, people, and objects) found on a remote island. For example, subjects were told to imagine that one member of the Barratos tribe was observed to be obese, and they estimated the proportion of all members of this group that would be obese. Likewise, the subjects were told that one sample of the substance "floridium" was observed to conduct electricity, and they estimated the proportion of all members of this set that would conduct electricity. There were several interesting findings from the Nisbett et al. study, but for our present purposes the most relevant is that the subjects were very sensitive to perceived variability of the conclusion category. For a variable category such as Barratos people (and their potential obesity), the subjects were rather unwilling to make strong inferences about other Barratos, after just one case. But for a homogenous category such as floridium samples, the subjects were willing to generalize the observation of electrical conductance to most or all of the population.

This result, that subjects were more willing to draw inferences about homogenous conclusion categories, makes a striking comparison to the results of Rips (1975). Whereas Rips found that typicality of the conclusion did not affect inductive strength, Nisbett et al.'s (1983) results show that conclusion categories do matter, at least in terms of their variability. The criteria for what makes a good premise category are different than the criteria for what makes a good conclusion category.

Studies with children on use of shared category membership. The Rips (1975) task has been adapted for testing with children, first by Carey (1985). There are a number of important reasons to study inductive reasoning in children. Such studies could show how inductive abilities develop, perhaps guiding or constraining accounts of fully developed, adult inductive reasoning. In comparing two models that equally account for adult data, if one model can also give an explanation of the course of development, then that model ought to be favored. Also, the performance of children on induction tasks can help the researcher determine what children know about a particular domain. For example, a pattern of age-related changes in reasoning about animals could reflect the growth of children's knowledge or theories about living things. Of course, with these different reasons for studying the development of induction, there is always the challenge of whether to attribute a change in performance to development of reasoning processes or development of knowledge.

Carey (1985) used an induction task with pictures of humans, animals, plants, and other things. Children, as young as age 4, were shown a picture of a premise item, such as a picture of a person, and told that it had some property, such as that of having a spleen inside.[1] Then the child was shown several pictures of other things, such as dogs, bees, and flowers, and was asked whether each also had the same property—for example, that of having a spleen. A number of results showed what makes a good case, from the point of view of young children. For children of age 6 and under, information about persons, as premises, tended to promote strong inductions. For example, when told that a person had a spleen, children were inclined to judge that dogs and bees had spleens as well. On the other hand, other animals did not make good cases, or were considered weak premises. For example, projection from dogs to humans was much weaker than projection from humans to dogs. This result maps well

onto the finding of typicality effects by Rips (1975), given the assumption that children consider humans to be very typical animals.

Some of the other results from Carey (1985) also map well onto past results, in particular that similarity effects were also found. A fact about humans was projected most strongly to other mammals, then to other animals such as birds and bees, and progressively less to plants and inanimate objects. Children as young as age 4 showed this pattern, but the steepness of the generalization gradient was greater in older children, suggesting greater sensitivity to similarity between categories.

On their own, these results from Carey (1985) would perhaps be described best as resulting from a change in knowledge rather than a change in processing. That is, as children get older, they lessen their use of humans as a prototype, and they increase the steepness of their generalization gradient when using information about similarity between various animals, plants, and inanimate objects. These changes could well reflect a maturing conception of things in the living world, and an increasing differentiation between various categories, rather than changes in how inductive reasoning is actually performed. However, this issue will be revisited as other results are described.

Chronologically, the Carey (1985) study was followed by several studies by Gelman and colleagues (Gelman, 1988; Gelman & Coley, 1990; Gelman & Markman, 1986; Gelman & O'Reilly, 1988;). In these experiments, similar procedures to Carey's were used, with children being told a property of some animal or object in a picture and then judging whether or not other animals or objects would also have this property. Gelman and Markman tested children as young as age 4, with the particular aim of looking at the nature of similarity effects. Although Carey did find similarity effects in young children, there are various ways in which animals and other things could be considered similar, and it is important to understand what kind of conception of similarity is guiding inductive inferences. Gelman and Markman contrasted similarity based on perceptual appearances with similarity based on underlying shared category membership. For example, a blackbird and a bat may look fairly similar, whereas a blackbird and a flamingo may not appear too similar, but the latter two share many internal characteristics because they are both birds. Using questions about unfamiliar internal properties, Gelman and Markman found that young children preferred to project between pairs of items with shared category membership, even when the members of the pair were less similar on the surface than those of some other pair. Therefore it was concluded that children used a fairly sophisticated conception of similarity to guide inductive reasoning, with deeper similarities such as category membership overriding more superficial similarities.

Gelman (1988) examined inductive reasoning at different levels of a taxonomic hierarchy (cf. Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). For example, given the premise that a daffodil has some novel property, children were asked whether the property would be true of another daffodil (same subordinate level category), a rose (same basic level category), a houseplant (same superordinate level category), and a bowl (unrelated category). Children as young as age 4 showed a generalization gradient that was similar to that for the results of Carey (1985), with the most projections within the subordinate level, and decreasing projections at higher levels of the taxonomy. Gelman pointed out that the results could reflect not only similarity-based reasoning but also sensitivity to category variability or homogeneity (as in Nisbett et al., 1983), with lower level categories being more homogenous. Indeed in some cases adult judgments of category homogeneity were significantly correlated with likelihood of children's inferences to that category. In a second study done by Gelman (1988), children were asked to generate their own familiar properties for the premise category and were then asked to judge whether this property held for the various conclusion categories as well. It was found that judgments about familiar properties followed the same patterns as did judgments about novel properties, suggesting that subjects might make inferences about novel properties by considering the distribution of known properties (see Heit, 1998).

As in Gelman's (1988) experiments, Gelman and O'Reilly (1988) looked at inductive reasoning at different points in a taxonomic hierarchy, but the focus was at higher or more superordinate levels of the taxonomy. Again, the pattern of results consisted in a decreasing likelihood of inferences at progressively higher taxonomic levels. This study also showed that children were equally willing to make inferences to typical superordinate category members and atypical superordinate category members, as in Rips (1975).

Gelman and Coley (1990) tried to extend the main results from older children to 2-year-olds, using a somewhat simpler procedure. The key findings were that children were able to use shared category membership to guide inferences, even when it conflicted with surface similarity, and that, again, the typicality of conclusion categories did not affect induction. In a useful control condition, it was found that the sharing of a category label, such as "bird," was critical to obtaining these results. Shared transient properties, such as "wet," did not serve as a reliable basis for induction.

At this point, it is worth mentioning parenthetically that despite this evidence for shared category membership as a basis for induction, in preference to perceptual similarity and other shared attributes, there is indeed evidence that perceptual similarity and other shared attributes do have some role in promoting inductive reasoning. That is, if two things look alike, you may still want to project a property from one to another on this basis, despite different category membership. In studies with children (as young as age 3) and adults, Florian (1994) found effects of perceptual similarity and shared attributes on induction, beyond effects of category membership. (See also Loose & Mareschal, 1999.)

Keeping with the progression of finding category membership as a basis for induction at increasingly younger ages, it is interesting to consider the relevant studies done with infants. Children as young as 3 or 4 years have shown a rather sophisticated use of similarity and category membership to guide induction. How well does this ability extend to even younger children? Again, induction tasks with infants could reflect their inductive reasoning abilities as well as their knowledge base for a particular domain of categories. Baldwin, Markman, and Melartin (1993) examined inductive inference in infants between 9 and 14 months of age, using an exploratory play task. The children learned, after a brief exposure of 30 sec, about a property of a novel toy—for example, that a can would make a wailing sound when squeezed. The test was whether children would expect another toy of similar appearance to have this property as well. Children played for a longer time with a second object that did not have the target property (e.g., a can that did not wail) in comparison with appropriate control conditions, suggesting that the infants had inferred that the second toy would be like the first toy and were surprised that it did not have the same property. This study showed that some of the inductive ability in older children is also present with infants, but it was not designed to look at whether or not children use a taxonomy of categories to support induction.

Mandler and McDonough (1996) also looked at induction with infants, but focused more on use of established taxonomies of categories. They looked especially at the use of the superordinates *animal* and *vehicle*. The infants were taught an action to perform on an object, such as giving a (toy) dog a drink; then they were tested on whether this action was generalized to other objects in the same superordinate category (e.g., a rabbit) and the other superordinate category (e.g., a bus). The children's pattern of play respected the boundaries of superordinate categories, with actions taught on one animal being extended to other animals but not to other vehicles, and likewise actions taught on one vehicle extended to other vehicles but not animals. However, there was not evidence for much sensitivity to distinctions within the superordinate category. That is, infants did not project more between similar items (e.g., *dog* and *rabbit*) than between dissimilar items (e.g., *dog* and *fish*), as long as all items were in the same superordinate category. There are a number of ways to conceive of this result. As in Gelman and Markman (1986), it shows the primacy of shared (superordinate) category membership. Unlike Gelman (1988), Mandler and McDonough (1996) did not address shared subordinate or basic level category membership, so it is difficult to say whether inductions would have been even stronger with subordinate or basic level categories than with superordinate categories (but see Mandler & McDonough, in press). Would the inferences have been even stronger from one kind of dog to another? Finally, the dissimilar conclusion categories in Mandler and McDonough (1996) tended to be atypical of the superordinate category (e.g., fish are less typical animals than are rabbits), so the lack of distinctions among various conclusion categories within a particular superordinate category could be taken as a replication of the finding that conclusion typicality does not affect induction (Carey, 1985; Gelman & O'Reilly, 1988; Rips, 1975).

Perhaps the best way to summarize the Baldwin et al. (1993) and Mandler and McDonough (1996) studies is that they documented an important subset of the known results for older children, using ordinary infant behaviors (playing with toys) as a means of measuring induction rather than other forms of the task which would require explicit judgments.

Further phenomena involving typicality, similarity, and specificity. One of the advantages of studying induction in adults rather than children is that it is possible to present a greater number of problems in one session and potentially address a greater range of phenomena. Indeed, Osherson, Smith, Wilkie, López, and Shafir (1990) have made a substantial and influential contribution to the study of inductive reasoning by documenting a set of important phenomena (see also Osherson, Stern, Wilkie, Stob, & Smith, 1991). Several of these phenomena involve reasoning from just a single premise category, whereas the remainder involve multiple premises and will be described in a later section of this paper. Osherson et al.'s (1990) experiments involved giving subjects pairs of inductive arguments such as the following:

Robins use serotonin as a neurotransmitter

Sparrows use serotonin as a neurotransmitter

and

Robins use serotonin as a neurotransmitter

Geese use serotonin as a neurotransmitter.

Subjects would tend to choose the first argument as stronger than the second, illustrating the premise–conclusion similarity phenomenon (reported by Rips, 1975, and subsequently others). In addition, Osherson et al. (1990) documented the premise typicality effect, reported by Rips and others. As an extension of this result, Osherson et al. (1990) described the premise–conclusion asymmetry phenomenon; for example, the argument from robins to geese above would be stronger than the reversed argument, from geese to robins (see also Carey, 1985, for an example of asymmetry). This phenomenon follows from the premise typicality effect, because whenever an argument has a premise category more typical than its conclusion category, the reversed argument should be weaker than the original argument.

Next, Osherson et al. (1990) documented the conclusion specificity phenomenon. It was found that arguments with a more specific conclusion category, such as *bird*, were considered stronger than arguments with a more general conclusion category, such as *animal*. This result makes sense from a logical perspective, in that more

evidence should be needed to support a more sweeping conclusion about a relatively superordinate category, in comparison with a narrow conclusion about a more subordinate category. (See also McDonald, Samuels, & Rispoli, 1996, for correlational evidence that over a range of arguments, scope of conclusion category is one of three good predictors of inductive strength.) This phenomenon can also be tied to the Nisbett et al. (1983) result showing that people make stronger inferences about more homogenous categories. In general, superordinate categories should be more variable than their subordinate categories, because the superordinate includes its own subordinates.

One of the important contributions of Osherson et al. (1990) was that they began to show where people's inductive inferences diverge from normative patterns. Rather than take the Humean approach of addressing whether induction can be justified, Osherson et al. (1990) aimed to show examples in which people's inductive inferences were clearly not justified—for example, because they violated axioms of probability. One relevant phenomenon is the inclusion fallacy, illustrated in the following arguments:

Robins secrete uric acid crystals

All birds secrete uric acid crystals

and

Robins secrete uric acid crystals

Ostriches secrete uric acid crystals.

People choose the first argument as stronger than the second, even though the first conclusion logically implies the second. Because the second conclusion is implied by the first conclusion, the probability of the second conclusion should not be lower than the probability of the first conclusion. The inclusion fallacy seems to reflect the use of similarity between the premise and conclusion in making judgments. If the representations of *robin* and *bird* are quite similar (due to the typicality of *robin*), but *ostrich* has a quite different representation than *robin*, then the second argument could seem weaker than the first. However, it should be noted that the more general result is the conclusion specificity phenomenon, in which arguments with more specific conclusion categories are considered stronger. The inclusion fallacy would seem to apply only in cases such as the example above involving a pair of category members, one very typical, such as *robin*, and one atypical, such as *ostrich*.

Also, using picture versions of the Osherson et al. (1990) tasks, López, Gelman, Gutheil, and Smith (1992) found evidence for a number of Osherson et al.'s (1990) single-premise phenomena with children ranging from ages 5 to 9. In addition to typicality effects, López et al. (1992) found conclusion specificity effects; for example, inferences were weaker for a conclusion about animals and plants than for a conclusion about just animals. This result, again, can be taken as support for Nisbett et al.'s (1983) findings of poor generalization to variable categories, here with young children. Likewise, the López et al. (1992) result converges nicely with the Gelman (1988) and Gelman and O'Reilly (1988) results that children draw strong inferences between items that are both members of a relatively specific, subordinate category.

Furthermore, Sloman (1993, 1998) extended the findings of the inclusion fallacy by demonstrating another phenomenon, inclusion similarity. As in the inclusion fallacy, the inclusion similarity phenomenon shows an effect of similarity between premise and conclusion categories, in an apparently nonnormative way. However, Sloman reported these effects even more dramatically, for deductively valid, perfectly strong arguments. Subjects found arguments of the form Animals/Mammals (i.e., with *animal* as the premise category and *mammal* as the conclusion category) stronger than arguments like Animals/Reptiles. Notably, both of these arguments are equally, and perfectly, valid. That is, anything true of all animals must necessarily be true of all mammals and all reptiles as well. However, an argument such as Animals/Reptiles may get a low strength rating because of relatively low similarity between the respective representations of *animal* and *reptile*.

Sloman (1998) documented a related phenomenon, called *premise specificity*, which also shows compellingly the influences of similarity on people's evaluations of inductive arguments. This phenomenon is well illustrated by the following example: People will prefer an argument with the form Birds/Sparrows over an argument with the form Animals/Sparrows. As in the inclusion similarity phenomenon, each argument is perfectly valid—there is no difference in the probability of the conclusion for one argument versus the other. Still, subjects describe the first argument, with a narrow premise category, as being stronger than the second, with a broad premise category. Sloman (1998) reported that the inclusion similarity and premise specificity findings were fairly robust over variations in procedure, but it is possible to prevent subjects from drawing such fallacious inferences by making the category inclusion relations explicit—for example, by reminding subjects that all sparrows are animals. It is also interesting to compare the premise specificity phenomenon with the conclusion specificity phenomenon (Osherson et al., 1990), in which people draw stronger inferences about a narrower conclusion, such as *bird* in comparison with *animal*. In contrast, conclusion specificity does seem to be compatible with axioms of probability theory.

Effects of expertise on induction. Developmental research on induction is important because, potentially, both knowledge and cognitive capacities are changing as children get older, allowing researchers to collect a very rich set of data. However, as mentioned, it can be difficult to attribute a developmental change uniquely to a change in knowledge or a change in cognitive mechanism. At the other end of the developmental continuum, adults are acquiring expertise on various topics as suited to their living conditions or working needs. Although adults with

different areas of expertise, or from different cultures, could possibly differ in terms of cognitive processing, it is plausible to attribute expertise differences in inductive reasoning largely to differences in knowledge. A recent, exciting trend consists of research on experts' inductive reasoning, going beyond past studies which had mainly looked at reasoning in American college students and American children. Coley, Medin, and Atran (1997) and López, Atran, Coley, Medin, and Smith (1997) studied inductive reasoning by Itzaj Mayans in the rainforest of Guatemala, people with great expertise regarding local plants and animals. Medin, Lynch, Coley, and Atran (1997) looked at inductive reasoning about categories of plants, by various kinds of tree experts.

Coley et al. (1997) looked at inductive reasoning at different levels of the taxonomic hierarchy of animals and plants. The purpose of this work was to see whether some taxonomic level is "privileged" or specially favored for inductive inferences, and whether this privileged level varies on the basis of expertise. For example, subjects were told to assume that all black vultures (a subspecies) are susceptible to a particular disease and were asked the likelihood that all vultures (a species, or strictly speaking a folk-generic category) would be susceptible to this disease. Coley et al. (1997) tested various premise–conclusion pairs, including the pairings of varietal and subspecies, subspecies and species, subspecies and lifeform (e.g., birds), species and life-form, and lifeform and kingdom (e.g., animals). The key result, for both Itzaj and American college students, was that there was a sharp drop in strength of inferences when conclusion categories were beyond the species level. That is, inferences regarding subspecies and species conclusion categories were quite strong, whereas inferences regarding the lifeform and kingdom level were much weaker. Coley et al. (1997) interpreted this result as showing that the species level is privileged, in that it is the broadest taxonomic level that supports strong inferences. However, specificity of the premise category did not seem to have much effect on induction, to the extent that Coley et al. (1997) compared premise categories at different taxonomic levels.

This study can be tied to a number of past results. First, the general result of weaker inferences for broader or more variable conclusion categories recapitulates the findings of Nisbett et al. (1983) and more recently the conclusion specificity phenomenon in Osherson et al. (1990) and López et al. (1992). Also there is some similarity between Coley et al.'s (1997) study and the work done by Carey as well as Gelman and colleagues, showing weaker inferences to the extent that the category encompassing the premise and conclusion items is more general or superordinate. However, the specific finding of Coley et al. (1997), that the species level is privileged, would not necessarily be predicted on the basis of past work. Indeed, it is surprising that the same level of privileged inference was found for the Itzaj and American college students, considering the knowledge differences—the far greater daily experience of plants and animals among the Itzaj. Coley et al. (1997) suggested that beliefs about categories' usefulness for induction could go beyond actually known facts and experiences. For example, American college students could simply have a belief that different species of animals have their own characteristic anatomies, diseases, and so forth, without any more specific knowledge to this effect (see also Heit, 1998, Shipley, 1993). So someone could treat a particular level as being privileged without detailed knowledge to support this distinction.

López et al. (1997) also compared induction by the Itzaj and by American college students and, in contrast to Coley et al. (1997), found more widespread influences of knowledge on patterns of inductive reasoning. They examined similarity and typicality effects, and found that the patterns of inductions differed between the two cultures in cases where their category representations diverged. For example, the Itzaj reported foxes as being more similar to cats than to dogs, whereas American students stated that foxes are more similar to dogs. This pattern was reflected by choices in a task where subjects saw pairs of inductive arguments. Itzaj subjects stated that arguments of the form Foxes/Cats were stronger, whereas Americans stated that Foxes/Dogs was stronger. Although Coley et al. (1997) did not find cross-cultural differences in the privileged level of conclusion categories, it is clear from López et al. (1997) that indeed there are some cultural, or knowledge-derived, differences.

Further evidence for effects of knowledge on induction comes from Medin et al. (1997), who looked at inferences about plant categories for three kinds of (American) tree experts: taxonomists, landscapers, and tree maintenance workers. Medin et al. were mainly interested in effects of similarity or shared category membership, for groups that differed among themselves regarding preferred taxonomic membership and upon occasions differed with regard to standard scientific taxonomies. For example, on a free sorting task, landscapers and maintenance workers tended to organize tree species in terms of their shape or utility for various landscaping tasks. Medin et al. (1997) devised questions on a test of inductive reasoning that pitted scientific matches against alternative, functional category structures. For example, two tree species might be distant in terms of the scientific taxonomy but they could both be useful for providing shade. The test items for the inductive inferences used biological properties concerning reproduction, disease, or physiology. It was found that taxonomists (not surprisingly) sorted trees on the basis of scientific taxonomy and likewise favored inductive arguments between categories that were close in the scientific taxonomy. Maintenance workers seemed to favor a more functional category organization for both sorting and reasoning. Landscapers seemed to be more flexible and possibly more conversant with multiple category structures; they tended to prefer functional organization for sorting but their biological inferences reflected knowledge of the scientific taxonomy. In sum, these three

groups of experts generally showed the similarity effects that have been documented in other populations, but the groups' knowledge about trees mediated these similarity effects. (See also Proffitt, Coley, and Medin, 2000.)

Discussion. The study of inductive arguments with a single premise category, has produced a number of interesting and consistent results. What promotes an inference from one case to another? The three key factors that promote inductive inferences are similarity between premise and conclusion category (in terms of a taxonomic hierarchy rather than superficial similarities), typicality of the premise category, and homogeneity of the conclusion category. Returning to the example at the start of this paper, how would these variables affect inferences about home burglaries? Similar cases should promote inference, so a burglary that is particularly near your home, or a recent burglary as opposed to one from the distant past, should increase the perceived risk for your own home. In addition, typicality of the given case should affect inferences, beyond any given effect of similarity. For example, if your neighborhood consists mainly of houses, a burglary in a houseboat should not generalize well to burglaries in other homes. The relation between houses and houseboats could well be asymmetrical, reflecting this difference in typicality. That is, characteristics of houses may seem to generalize to characteristics of houseboats better than characteristics of houseboats would generalize to those of houses. Finally, variability of the conclusion category should lead to weaker inferences. For example, a home that is sequentially occupied by different people with different habits and different possessions and sometimes is completely unoccupied would be relatively difficult to make predictions about, compared with a more stable conclusion item.

These three results have ties to past philosophical work on induction. Similarity effects, or the idea that seeing some commonalties between two items should promote the inference of further commonalties, has been a longstanding position in philosophy (see, e.g., Mill, 1874). However, Goodman (1972) has argued that similarity itself may not be a primitive notion; for example, the features that are used to assess similarity can be context dependent (see Hahn & Chater, 1997, and Medin, Goldstone, & Gentner, 1993, for reviews of related psychological results). Likewise, Gelman and Markman (1986) made an important distinction between inductions based on internal similarity and those based on external, perceptual similarity. This issue will be returned to in the third section, which describes results in which the use of similarity depends on the property being projected.

Furthermore, Shipley (1993) has applied Goodman's (1955) work on induction to several results in psychology. The idea that some categories, such as more typical categories, are particularly good for promoting inferences, in part because of their past frequency of use, is related to Goodman's (1955) idea of entrenchment of predicates, with some predicates (or categories) promoting inferences more than others. However, this analysis does not explain why entrenchment, or typicality, of premise categories matters but typicality of conclusion categories does not. And again, the point could be made that typicality is not a primitive concept any more than similarity, and there could be several determinants of typicality such as frequency of use, centrality, similarity to other category members, and nearness to an ideal (cf. Barsalou, 1985).

Finally, the use of beliefs about variability of conclusion categories, or beliefs that some taxonomic level may be privileged for induction, is tied to Goodman's (1955) concept of overhypotheses. Overhypotheses are general beliefs that guide inference, without necessarily having much specific content. For example, a person can believe that samples of a particular kind of metal will be homogenous in terms of whether they conduct electricity, without any more specific knowledge of whether this kind of metal does in fact conduct electricity. The sensitivity to conclusion variability in Nisbett et al. (1983) can be explained in terms of use of overhypotheses about different kinds of categories (metals, people, etc.).

To conclude, it is interesting to note the results that have not been reported. For example, there have been no reports to date of independent effects of the typicality of the conclusion category as opposed to the premise category. Indeed, it would seem useful for any account of induction to address why conclusion typicality does *not* matter, even as it explains why premise typicality does matter. Another nonresult relates to the homogeneity or variability of a premise category. It is clear that conclusion homogeneity promotes inferences, but no studies have directly addressed the effects of homogeneity of a single premise category. (Sloman's, 1998, premise specificity effect comes the closest, but this result seems to have depended heavily on similarity between the premise and conclusion categories, rather than on the homogeneity of the premise category.) This issue may be somewhat easier to study with multiple premise categories, because the diversity within, say, a pair of premise categories can be manipulated easily by choosing similar versus different pairings. The next section will address premise diversity as well as several other important results that have been obtained by studying induction with multiple categories.

What makes a good set of cases?

When people try to make an inference about some object or event, they are typically faced with a great deal of information. Rather than just one past case being available or relevant, in many realistic situations there will be an extensive set of cases that could be relied on. How do people draw inductive inferences from multiple cases? What makes a set of cases or precedents seem strong, compelling, or useful for promoting inferences? One factor is numerosity. For example, the more homes that have been broken into on your street, the greater the perceived risk for your own home. However, one of the fascinating characteristics of human inductive inference is that people do not simply add up evidence from individual cases. That is, putting together two cases that are strong on their

own does not necessarily lead to an even stronger argument based on both cases. In the first part of this section, the evidence for when numerosity does increase inductive strength will be covered, then the evidence for more complex and subtle phenomena, dependent on the diversity and variability of cases rather than their numerosity, will be reviewed.

Number of cases. In their study involving inferences about people and objects on an island, Nisbett et al. (1983) systematically varied the given number of observations. For example, subjects were told that 1, 3, or 20 obese members of the Barratos group had been observed and were asked what proportion of all Barratos are obese. In general, inferences were stronger with increased sample size. However, this effect interacted with homogeneity of the conclusion category. If the conclusion category was perceived as very homogenous (e.g., floridium samples with respect to electrical conductivity), then just one case was enough for subjects to generalize to the whole population (or nearly 100%). Therefore there was something of a ceiling effect, and increases in sample size did not always lead to higher estimates.

Osherson et al. (1990) referred to the sample size effect as *premise monotonicity*—namely, a monotonic relation between the number of premise categories in an inductive argument and rated inductive strength. Although they found interesting exceptions to this phenomenon, to be described shortly, the overall trend supported this generalization. Likewise McDonald et al. (1996) measured inductive strength for a variety of arguments and found that the number of premise categories in the argument was one of the reliable predictors of strength.

Not only does sample size or number of premise categories serve as a robust determinant of inductive strength in adults, but in some cases children's inductive inferences appear to be sensitive to sample size. In particular, both López et al. (1992) and Gutheil and Gelman (1997) found some evidence for sample size effects in 9-year-olds. López et al. (1992) used a picture version of the Osherson et al. (1990) task, and found that 9-year-olds favored an argument of the form Raccoon, Leopard, Skunk, Tiger, Giraffe/Animal over the form Skunk, Tiger, Giraffe/Animal. More premise categories led to greater inductive strength. However, the sample size effect was not entirely robust, even in 9-year-olds. Children of this age did not show sample size effects for similar arguments with a more specific conclusion category—that is, *bear* rather than *animal*. López et al. (1992) interpreted this difference between more general and more specific conclusion categories in terms of the account proposed by Osherson et al (1990). According to Osherson et al. (1990), evaluating an argument with a specific conclusion category such as *bear* would require the generation of a superordinate category, such as *animal* or *mammal*. Therefore the arguments with a specific conclusion would require more cognitive processing and hence would be more difficult overall, masking any sensitivity to sample size. In contrast, López et al. (1992) failed to find any sensitivity to sample size among 5-year-olds for both general and specific arguments, even in a task in which the experimenter counted the number of premise categories for the child. At present, there seems to be no evidence that children younger than 9 use sample size evidence in inductive reasoning, although it is tempting to imagine that sample size is such a central element of reasoning that in the future procedures might be devised to find sensitivity in younger children.

Gutheil and Gelman (1997) also looked at sample size effects in 9-year-olds. As in López et al. (1992), there was actually mixed evidence, with the children failing to show sensitivity to sample size in some cases. Gutheil and Gelman used a similar procedure to that of López et al. (1992), describing hidden properties of animals, but with categories at a somewhat lower taxonomic level. All of the premise items were in the same basic level category (e.g., they were all frogs). On the basis of past work (e.g., Gelman & O'Reilly, 1988) showing that children's inferences are stronger at lower taxonomic levels, it was hoped that the sample size effect would be more evident at lower levels. Gutheil and Gelman used a specific conclusion (e.g., a picture of another frog), and in their first attempt they did not find sample size effects in 9-year-olds, essentially replicating López et al. (1992). In a second study, however, they simplified the task by not showing the picture of the conclusion item but simply describing it. Here, Gutheil and Gelman found sample size effects—namely, stronger inferences based on five premise items as opposed to one premise item.

Diversity of cases. Although sheer numerosity of cases does have some effect on induction, there is also substantial evidence that variability or diversity of cases affects inductive strength. Intuitively, repeating the same evidence, or highly similar pieces of evidence, again and again should not be much more convincing that just giving the evidence once. On the other hand, if different kinds of converging evidence come from different sources, then potentially a stronger or broader case can be made. This result, that more variable observations promote broader or stronger generalizations, is now considered a truism in areas of research near to induction, such as categorization (e.g., Fried & Holyoak, 1984; Homa & Vosburgh, 1976; Posner & Keele, 1968).

The first study of diversity-based reasoning in induction was a developmental one by Carey (1985), comparing 6-year-olds and adults. Carey looked at patterns of inductive projection, given the premises that two diverse animals, dogs and bees, have some biological property. The purpose of this study was to see whether subjects would reason that "if two such disparate animals as dogs and bees" had this property then "all complex animals must" (p. 141). Indeed, adults made broad inferences to all animals, extending the property not only to things that were close to the premises (other mammals and insects) but also to other members of the *animal* category (such as birds and worms). In contrast, the children seemed to treat each premise separately; they drew inferences to

close matches such as other mammals and insects, but they did not use the diversity information to draw a more general conclusion about animals. Therefore, in this first attempt there was evidence for effects of diversity in adults but not children. However, Carey was simultaneously interested in development of reasoning as well as development of the *animal* concept. The nonappearance of the diversity effect in children could have been due to an undeveloped *animal* concept in 6-year-olds, rather than different or incomplete processing.

In a follow-up study, Carey (1985) looked at diversity effects based on the concept *living thing* rather than *animal*. The most relevant result was that subjects were taught a biological fact either about dogs and bees or about dogs and flowers, with the latter being even more diverse than the former. Given a fact about dogs and flowers, children did tend to generalize fairly broadly, suggesting that children may have some sensitivity to diversity of premise categories. However, if anything, they tended to overgeneralize, extending the property not only to other living things but often to inanimate objects as well. Therefore Carey concluded that 6-year-old children did not quite have a developed *living thing* concept serving as the basis for induction. Still, there was suggestive evidence for the impact of diversity of premise categories in this study.

Continuing along this line of research that looks for diversity effects in children, López et al. (1992) found limited evidence for 9-year-olds and no evidence for 5-year-olds. For the 5-year-olds, choices in a picture-based task did not show any sensitivity to diversity of premise categories, even when the diversity was explicitly mentioned by the experimenter. However, 9-year-olds did show sensitivity to diversity of premises, but only for arguments with a general conclusion category such as *animal* rather than a specific conclusion category such as *kangaroo*. Again, López et al. (1992) explained this result in terms of arguments with specific conclusion categories' requiring more stages of cognitive processing than are needed for arguments with general conclusion categories.

Gutheil and Gelman (1997) attempted to find evidence of diversity-based reasoning for specific conclusions in 9-year-olds, using category members at lower taxonomic levels which would presumably enhance reasoning. However, like López et al. (1992), Gutheil and Gelman did not find diversity effects in 9-year-olds, although in a control condition with adults, there was robust evidence for diversity effects.

More recently, however, Heit and Hahn (1999) reported diversity effects in children younger than 9 years, in experiments with pictures of people and everyday objects as stimuli rather than animals with hidden properties. For example, children were shown a diverse set of dolls (a china doll, a stuffed doll, and a Cabbage Patch doll), all being played with by a girl named Jane. Also children were shown a nondiverse set: three pictures of Barbie dolls, being played with by Danielle. The critical test item was another kind of doll, a baby doll, and the question was, Who would like to play with this doll? In another stimulus condition, there was a diverse set of hats worn by one person, and a nondiverse set worn by another person, and again, the critical question was whether another hat would belong to the person with diverse hats or the person with nondiverse hats. For 74% of these critical test items, children 5 to 8 years of age made the diverse choice rather than the nondiverse choice. It seems from the Heit and Hahn experiments that children can follow the diversity principle at some level. However, it will take further work to establish the critical differences that led the past studies not to find diversity effects in children.

Indeed for adults, or at least American college students, there has been considerable evidence for diversity-based reasoning. Osherson et al. (1990) documented diversity effects in adults, for written arguments with general as well as specific conclusion categories. López (1995) devised a stricter test of diversity-based reasoning, in which people chose premise categories rather than simply evaluate arguments given a set of premises. In other words, would people's choices of premises reveal that they valued diverse evidence? Subjects were given a fact about one mammal category, and they were asked to evaluate whether all mammals had this property. In aid of this task, the subjects were allowed to test one other category of mammals. For example, subjects would be told that lions had some property, and then they were asked whether they would test leopards or goats as well. The result was that subjects consistently preferred to test the more dissimilar item (e.g., goats rather than leopards). It appears on the basis of López that for inductive arguments about animals, subjects do make robust use of diversity in not only evaluating evidence but also seeking evidence. (See also Spellman, López, & Smith, 1999, for a comparison with other reasoning tasks involving evidence selection.)

Do adults in other cultures show evidence of diversity-based reasoning? One might think that, just as diversity effects are age-dependent, they might also depend on knowledge or cultural experience. Choi, Nisbett, and Smith (1997) reported diversity effects in Korean university students, for both animal categories and categories of people. However, in their study of Itzaj adults in Guatemala, López et al. (1997) did not find evidence for diversity-based reasoning, using arguments with various categories of living things and questions about disease transmission. Indeed, sometimes Itzaj subjects reliably chose arguments with homogenous premise categories over arguments with diverse categories. (See also Coley, Medin, Proffitt, Lynch, & Atran, 1999.) From the subjects' explanations, it seems that they were using other knowledge about disease transmission that conflicted with diversity-based reasoning. For example, given a nondiverse argument, that two similar kinds of tall palm trees could get a certain disease, one subject claimed that it would be easy for the shorter kinds of palm trees, below, to get the disease as well. This issue, of how knowledge about properties guides induction beyond the structural effects of the categories themselves, will be discussed ex-

tensively in the next section on what makes a good property for induction. It does appear that the appearance of diversity may depend on relevant supporting knowledge's being accessed. In a follow-up study, López et al. (1997) found that the Itzaj did show diversity-based reasoning effects in some contexts. For example, Itzaj subjects were told to imagine buying several bags of corn. The question was whether it would be better to inspect two corn cobs from one bag, or one corn cob from each of two different bags. (See Nagel, 1939, p. 72, for a related example.) The subjects tended to prefer the latter, more diverse choice. This important result suggests, following Carey (1985), that diversity-based reasoning depends not only on processing but on knowledge.

Exceptions to diversity effects. The lack of diversity effects found in the Itzaj people suggests that there may well be other systematic responses to diverse information, and that in some cases diverse premise categories may not lead to a very convincing argument. In their influential work, Osherson et al. (1990) documented situations in which more diverse premise categories actually led to weaker inferences, referring to these as *nonmonotonicity effects*. For example, consider the following pair of arguments:

Flies require trace amounts of magnesium for reproduction

Bees require trace amounts of magnesium for reproduction

and

Flies require trace amounts of magnesium for reproduction

Orangutans require trace amounts of magnesium for reproduction

---

Bees require trace amounts of magnesium for reproduction.

Adult subjects tended to judge the first argument as stronger than the second, in apparent contradiction to both the sample size and diversity phenomena. According to Osherson et al. (1990), the reason why the second argument seems weaker is that it brings to mind a broader superordinate context, animals rather than insects. Whereas flies are highly typical insects, in the context of animals, flies are much less typical and orangutans are not prototypical either. Hence the second argument would be weaker because of lower typicality of premise categories.

Sloman (1993) reported a related violation of diversity, referred to as the *feature exclusion effect*. It was found that most subjects found an argument of the form Foxes, Deer/Weasels to be stronger than an argument of the form Foxes, Rhinos/Weasels, despite the greater diversity of the latter set of premises. According to Sloman (1993), the reason for this result was that rhinos and weasels have so few features in common (i.e., they are so dissimilar) that adding information about rhinos to a statement about foxes just does not warrant any further conclusions about weasels.

Finally, it is useful to mention that this nonmonotonicity effect has been replicated with Korean undergraduates (Choi et al., 1997), and furthermore that López et al. (1992) found some evidence for nonmonotonicity effects in 5-year-olds and even more consistent evidence with 9-year-olds. Perhaps the best conclusion to be drawn from nonmonotonicity effects as well as feature exclusion effects is that although diverse premises promote induction, too much diversity can actually hurt rather than help.

Discussion. When multiple premises are used to evaluate an inductive argument, the associated phenomena are rather interesting and varied. The key results can be summarized in terms of two main findings as well as the exceptions to these findings. The main findings are, again, that higher numbers of premise categories, as well as diversity of premise categories, promote inferences. The sample size effect seems to be robust, although its empirical status could be clarified for children of age 9 and younger. The diversity effect seems to be less robust, in that there are cultural or knowledge-based differences as well as a number of negative results with children. Some of the negative findings with respect to sample size and diversity seem consistent enough to treat as phenomena in their own right—for example, nonmonotonicity effects (Osherson et al., 1990) and feature exclusion effects (Sloman, 1993).

Because the results are particularly variable for diversity effects, it is useful to systematically enumerate why in a particular situation a person, whether child or adult, may not show diversity-based reasoning (and see Coley et al., 1999, for a further discussion). This question is especially interesting, considering that it seems normative to draw stronger inferences from more diverse observations. This claim has been made by philosophers such as Nagel (1939) in the context of probability theory and Hempel (1966) in the context of scientific inference from experiments (see also Bacon, 1620/1898; Heit, 1998; López, 1995). Note that the point of these claims was not to provide a complete justification for inductive inference, but rather to argue that diverse evidence may be more likely to satisfy particular goals. For example, Hempel claimed that conducting diverse experiments is compatible with a falsifying strategy in testing a scientific theory, compared with conducting a series of similar experiments.

One class of explanation for a lack of diversity effects, say in children, would consist of processing differences. For example, López et al. (1992) suggested, following the model of Osherson et al. (1990), that adults carry out a two-stage procedure in assessing inductive strength, assessing premise-to-conclusion similarity as well as the diversity of the premise categories (or how well they cover a generated superordinate). The lack of diversity effects in children could be due to an abbreviated procedure in which they complete the first stage but not the second.

Processing explanations bring up the question of whether processing in children is truly different from adult processing, or simply more fragile. Perhaps under the right conditions—for example, with simple materials that minimize task demands—children could show the same processing as do adults.

Another class of explanation comprises knowledge differences, which was highlighted by the cross-cultural studies of López et al. (1997), who showed domain differences in diversity for Itzaj adults in Guatemala. Likewise, Carey (1985) treated the diversity task as a measure of the maturity of various concepts such as *living thing* and *animal*.

Finally, it is possible that when a group of subjects, say children, fail to show diversity effects, they do so because there is a mixture of systematic responses. For example, about half of the time the children might be showing diversity effects, whereas for various reasons the other half of the time they could be doing something else systematically, such as being affected by the feature exclusion effect. Indeed, there are borderline results in both López et al. (1992) and Gutheil and Gelman (1997) that are opposite to diversity, suggesting the possibility that children were systematically doing something different rather than simply guessing. For a particular nonfinding of diversity, explanations due to missing processing mechanisms, performance difficulties, knowledge effects, or other systematic effects would all be possible. It is suggested that a future goal of studies on diversity should be not only to document when diversity-based reasoning does and does not occur, but to specifically aim at distinguishing among different explanations for its nonoccurrence.

As mentioned in the section on single-premise arguments, the studies of premise diversity effects facilitate the comparison with studies on conclusion variability (e.g., Nisbett et al., 1983). Whereas having a variable or broad conclusion category leads to weaker inferences, it now seems that having a variable or broad set of premise categories generally leads to stronger inferences (at least in American adults). It seems that people are concerned about breadth of categories for both premises and conclusions of inductive arguments, but breadth of premises leads to the opposite result of breadth of the conclusion category. It would be very interesting for models of induction to address directly why this variable has different effects for premise categories and conclusion categories.

In conclusion, to return to the example of homes and burglaries, it is useful to consider what would make a set of cases likely to promote inferences about another burglary. The effect of sample size has an intuitive effect; the more burglaries on your street, the higher your perceived risk. Can diversity effects be tied to the home burglary example? Perhaps. Say that it is the first of February. If there were a dozen burglaries on your street last year, one in each month, that may seem to indicate a fair risk for your own home. On the other hand, what if there were a dozen burglaries, all taking place on Christmas eve? This situation would involve more recent events, but they seem to form a localized or restricted cluster. The other situation, with a greater diversity of burglary occasions over the span of a year, all more distant from the present date, might promote a stronger inference about the present situation.

What makes a good property?

So far, this review has focused on the effects of categories on induction; that is, what makes a set of categories promote inductive inference? This emphasis has followed the historical emphasis of the field; for example, three of the most influential studies of induction (Carey, 1985; Osherson et al., 1990; Rips, 1975) also focused on categories. However, properties or predicates also have a crucial role in inductive reasoning—the end part of a statement, such as *thrives in sunlight* or *secretes uric acid crystals*, has considerable effects on how people respond to inductive arguments. In the example of homes, it makes intuitive sense that different predicates will have different patterns of projection. For example, if your neighbor's home is burglarized, the perceived risk for your own home seems greater. The proximity between the two homes promotes this inference. However, if your neighbor's home is painted blue, that does not seem to increase the risk that your own home will be painted blue. For this predicate, proximity does not have much predictive value. In this section, several ways that properties matter will be reviewed. A number of past results on property effects can be described as relating to the scope of the property. For example, house color is a stable or consistent property for one house, but it tends to vary more within a group of nearby houses. Other property effects could be attributed to differing use of similarity information for different properties. In many past studies already reviewed, subjects have seemed to reason about biological properties of animals in terms of some notion of internal similarity—projecting, for example, more readily from horses to cows rather than to lizards. But for other properties, such as house burglaries, the relevant measure of similarity might be physical proximity. Finally, a number of other ways that the content of properties influences inductive reasoning will be reviewed.

Scope of properties. The Nisbett et al. (1983) study is a good first illustration of how knowledge about the scope of a property affects inductive inference. As already reviewed, seeing that just one member of the Barratos group is obese does not seem to promote the inference that other people in this group will be obese. Obesity seems to be more of an individual characteristic rather than a group characteristic. On the other hand, Nisbett et al. found that people make stronger inferences for the same category but another property, skin color. Here, seeing the skin color of just one Barratos promotes inferences about other members of this group, on the assumption that members of the same ethnic group will likely have some shared physical characteristics. In another study with adults, Gutheil and Gelman (1997) reported property effects

like those found by Nisbett et al., but for a wider range of properties. In the terminology of Goodman (1955), it appears that some properties are more projectable than others.

This use of knowledge about scope of properties is not limited to adults but is clearly evident in young children as well. For example, Gelman (1988) compared stable, internal properties with more transient or idiosyncratic properties, in reasoning tasks performed by children as young as age 4. For projectable properties such as *has pectin inside*, children's inferences showed similarity effects, reflecting the taxonomic hierarchy of categories. But for properties such as *has a little scratch on it*, children showed chance patterns of reasoning, indicating that for properties with an idiosyncratic scope, they did not have a systematic basis of projection. (Also, see Springer, 1992, for similar results, in which children used kinship information to project biological properties, but projected idiosyncratic properties at a chance level.)

Young children's reasoning about the scope of properties is surprisingly sophisticated. A study by Macario, Shipley, and Billman (1990) showed rather subtle use of information about property variability by 4-year-olds. In particular, children were able to use the variability of one property to infer the variability of another property. The task was to learn about groups of objects that were preferred by one puppet or another. For example, children would see that the objects in one group all were blue and that a contrast category had one red member. Then the children were presented with a set of transfer items for classification, and on the basis of these choices it appeared they had inferred that the contrast category's other members would all be red. Likewise, after seeing that one category's members varied in shape, children inferred that the contrast category's members would also vary in shape. As in Nisbett et al. (1983), it was demonstrated that children would more readily base their inferences on a homogenous property as opposed to a property that varied across category members. But in addition it was shown that children could infer the variability or scope of a property in a sensible and productive manner.

More recently, Waxman, Lynch, Casey, and Baer (1997) have looked at knowledge about scope of properties for real animal categories. In an initial experiment, Waxman et al. found that given a property of, say, a collie, young children tended to extend this property to other members of the same subordinate category (other collies) as well as other members of the same basic level category (other dogs). As in Macario et al. (1990), the question was whether children could learn about scope of properties, and in particular whether they would infer that some properties were distinctive for different subordinate categories, but homogenous within each subordinate. Children were taught facts about two subcategories, such as that one breed of dog was used to find birds and another breed of dog was used to pull sleds. Then the children were taught that a dog of a third kind had another characteristic, such as that of being used to help take care of sheep. Finally, the children were tested on whether this third characteristic would extend to a variety of dogs and other animals. Unlike in the initial experiment, when this training was provided the children tended to restrict the scope of their inferences to the original subordinate category. With a small amount of training, 4-year-old children were able to learn about the scope of a property and use this information in a consistent way.

It seems that even infants show evidence for increasing sophistication about the scope of properties. Mandler and McDonough (1998) used an imitation task to compare 14-month-olds and 20-month-olds on inferences with properties that would have a scope at the basic level of categorization (e.g., chewing on bones would apply to dogs but not other animals such as birds). It was found that the 14-month-olds were willing to project properties rather widely, such as projecting bone-chewing to birds, but that the 20-month-olds were more restricted in the breadth of their generalizations, suggesting that they were sensitive to the scope of these properties.

Properties and similarity. Although it might seem from the previous section that some properties have a wide scope for projection whereas other properties are simply idiosyncratic and harder to project, the picture is actually more complicated and more interesting. Depending on the argument—that is, depending on the categories in an inductive argument—a particular property may be projectable, nonprojectable, or somewhere in between. Consider the following example, from Heit and Rubinstein (1994). For a typical blank anatomical property, such as *has a liver with two chambers*, people will make stronger inferences from chickens to hawks than from tigers to hawks. Because chickens and hawks are from the same biological category and share many internal properties, people are quite willing to project a novel anatomical property from one bird to another. But since tigers and hawks differ in terms of many known internal biological properties, it seems less likely that a novel anatomical property will project from one to the other. This result illustrates the priority of biological categories that has been observed in induction (e.g., Carey, 1985; Gelman, 1988). However, now consider the behavioral property *prefers to feed at night*. Heit and Rubinstein found that inferences for behavioral properties concerning feeding and predation were weaker between the categories *chicken* and *hawk* than between the categories *tiger* and *hawk*—the opposite of the result for anatomical properties. Here, it seems that despite the considerable biological differences between tigers and hawks, people were influenced by the known similarities between these two animals in terms of predatory behavior, thus making strong inferences about a novel behavioral property. In comparison, chickens and hawks differ in terms of predatory behavior (with chickens tending to be pacifists), so that people were less willing to project a novel behavioral property between these two animals. Together, these results suggest that each property is more projectable for a different pair of animals. (Also see Choi et al., 1997, for a comparison between anatomical and behavioral properties.) It is not simply the case that some properties

are always more projectable than other properties. Instead, there was a crossover interaction pattern between properties and premise–conclusion matches.

Recently, Ross and Murphy (1999) have also provided evidence for the flexibility of people's reasoning about different kinds of properties. Ross and Murphy's interest was the domain of foods, which perhaps in comparison with other domains such as the animal kingdom leads more readily to cross classification. For example, a bagel can be considered as part of the breads category (a taxonomic organization) or as a breakfast food (a script based organization). (See also Murphy & Ross, 1999.) Ross and Murphy compared two kinds of properties: biochemical properties and situational properties relating to how a food might be used. It was found that for biochemical properties, subjects preferred inferences based on taxonomic matches, whereas for situational properties, subjects preferred script-based matches. Just as in Heit and Rubinstein (1994), any account of induction that does not take into account the property being projected could not account for these results. In particular, inductive inference cannot be reduced to simply assessing the similarity between premise and conclusion categories, unless a flexible conception of similarity is allowed, in which similarity depends on the property being projected. For example, inferences about behavioral or situational properties might lead to behavioral or situational features' being emphasized in similarity computations. Smith, Shafir, and Osherson (1993) referred to such an effect as *feature potentiation* (and see Heit, 1997, for a review of related work). However, this term in itself does not give an account of how the process would take place. It seems likely that feature potentiation would rely on other mechanisms of memory (of which inferences have been successful in the past) as well as explanatory reasoning (about which features might be useful).

There is also some evidence that this kind of property effect occurs in children's reasoning as well. Gelman and Markman (1986) provided children with a property for one item (e.g., a blackbird) and then asked them whether the property would be true for a perceptually similar item (e.g., a bat) or an item that was a taxonomic match but less similar perceptually (e.g., a flamingo). It was found that for biological properties (e.g., referring to eating habits), the children preferred the taxonomic match, but that for perceptual properties (e.g., texture), they were at chance level or in some cases they showed a tendency to choose the perceptual match, suggesting that different features might have been potentiated for perceptual inferences.

Using a somewhat different task, Kalish and Gelman (1992) have looked at property inferences based on novel combined categories, such as *glass scissors*. Children were given facts about one of these categories, such as *used for partitioning* (a functional property) or *will get fractured if put in really cold water* (a dispositional property). The subjects were asked whether these properties would be true of other items as well, such as metal scissors and a glass bottle. The children (age 4) preferred matches in terms of object kind (e.g., both scissors) when projecting a novel functional property, but they preferred matches in terms of composition (e.g., both glass) when projecting a novel dispositional property, showing an impressive degree of sophistication about inferences.

Moving to even younger ages, Mandler and McDonough (1996, 1998) reported sensitivity to different kinds of properties for infants. Using a task in which 14-month-old children imitated actions performed on various objects, they found that the children were sensitive to the difference between animal actions (e.g., giving a drink) and vehicle actions (e.g., opening with a key). The children were less likely to repeat animal actions performed on a vehicle or vehicle actions performed on an animal than they were to repeat actions that matched the items. Again, there was evidence that quite young children are sensitive to the idea that there are different kinds of properties, with differing relevant criteria for projecting these properties. At no point during the course of their development has it been demonstrated that children treat all properties as the same—the default seems to be to show property effects of some kind.

Other property effects. In addition to the property effects just reviewed, researchers have documented a number of other interesting phenomena deriving from the content of properties. The diversity of these phenomena attests to the importance and prevalence of property effects, touching on several aspects of inductive reasoning. What these phenomena have in common, however, is that they all point to the limitations of similarity as a basis for inductive inference. Smith et al. (1993; see also Osherson, Smith, Myers, Shafir, & Stob, 1994) provided an important example in which inferences go in the opposite direction of what overall similarity would predict. Consider the following two arguments:

Poodles can bite through barbed wire

German shepherds can bite through barbed wire

and

Dobermans can bite through barbed wire

German shepherds can bite through barbed wire.

Clearly there is greater similarity between Dobermans and German shepherds than there is between poodles and German shepherds. Yet people find the first argument stronger than the second. An informal way to justify this reasoning is that if poodles, a rather weak and tame kind of dog, can bite through barbed wire, then obviously German shepherds, which are much stronger and more ferocious, must be able to bite through barbed wire as well. This property, *can bite through barbed wire*, seems to depend on the magnitude of other dimensions such as strength and ferocity. Again, informally, it seems that subjects are trying to explain how the various animals could bite through barbed wire, in terms of known facts about these animals.

However, this result could be explained alternatively in terms of the diversity effect, on the assumption that in addition to the premises provided to subjects, people use

their own prior knowledge to create additional, hidden premises. For example, people might already believe that another large, ferocious kind of dog, such as a Rottweiler, can bite through wire. This belief could serve as a hidden premise that would affect judgments about the conclusion. In this situation, supplying the premise that poodles can bite though barbed wire would lead to a diverse range of premise categories, Rottweilers and poodles. In contrast, supplying the premise that Dobermans can bite through barbed wired would represent a fairly narrow set of premises, Rottweilers and Dobermans. Hence, following the already established diversity effect, the premise with poodles should lead to a stronger conclusion.

Sloman (1994, 1997) has investigated the role of explanations in inductive reasoning more directly. Sloman has concluded that people are highly sensitive to the content of properties being projected, coming up with an explanation of the manifested property as a means of assessing inductive strength. An argument will be strong to the extent that premise and conclusion statements have the same explanations. For example, consider the following.

Many ex-cons are hired as bodyguards

Many war veterans are hired as bodyguards

and

Many ex-cons are unemployed

Many war veterans are unemployed.

According to Sloman (1994), the first argument is strong because both statements have the same explanation—namely, that ex-convicts and war veterans are hired as bodyguards because in both cases they are tough and experienced fighters. The second argument is weaker because the two statements would have different explanations—namely, that ex-convicts might be unemployed for different reasons than war veterans. (Sloman, 1997, investigated this phenomenon further, distinguishing between unrelated explanations and conflicting explanations.) As in the Smith et al. (1993) results, and for that matter the results of Heit and Rubinstein (1994) and Ross and Murphy (1999), it seems that inductive inference with meaningful properties critically depends on determining which known characteristics of the categories are causally related to or predictive of the property to be projected. Indeed, when Lassaline (1996) made various causal relations explicit to subjects, she found that they were particularly sensitive to causal relations between characteristics of the premise category and the property to be projected. (See also Hadjichristidis, Sloman, Stevenson, & Over, 1999, and Wu & Gentner, 1998).

Finally, as possibly converging evidence for the role of explanations, McDonald et al. (1996) found that number of conclusions suggested by a set of premises was negatively correlated with perceived inductive strength. Perhaps when a set of premises all have the same explanation, it leads to a single, clear and strong conclusion, but when there are multiple conclusions it is reflective of conflicting possible explanations and thus any particular inference will be weak.

Discussion. The main conclusion from this section is that properties matter, a great deal! In addition to factors such as similarity between premise and conclusion categories, and typicality and diversity of premise categories, the content of the property being projected from premise to conclusion has a central role in inductive inference. Perhaps most dramatically, idiosyncratic properties such as being obese or having a scratch do not lead to widespread, systematic inferences. In addition, Smith et al. (1993) showed that for some properties, similarity between premise and conclusion categories is negatively correlated with inductive strength, although this result could also be explained in terms of diversity. To account for other results (e.g., those of Heit and Rubinstein, 1994), one must assume that different kinds of similarity would be used for inferences about different properties, fitting with Goodman's (1972) points about the flexibility of similarity. Sensitivity to different kinds of properties has been observed in young children and even infants. If one thing is clear, it is that any complete account of inductive reasoning needs to address property effects. Many valuable and systematic results, reviewed in the first two sections of this paper, have been obtained from studies in which properties were not varied systematically, but it seems that these studies were looking at only a restricted range of human abilities. Just as it is possible to learn more about the cognitive processes underlying induction by using arguments with multiple premises rather than a single premise, it is possible to learn yet more about induction by comparing performance with different properties.

In all three sections of this paper so far, a few themes have emerged repeatedly. One is that, as Goodman (1955) noted, categories and properties vary in terms of their entrenchment. Some categories and some properties seem to be more suitable for inductive reasoning than others. Although Goodman referred to this issue as a "riddle," it seems that humans are rather systematic in terms of what they treat as more or less entrenched; typical categories, for example, tend to be good for induction whereas transient properties tend to be bad for induction. Another theme is that for categories as well as properties, there is a sense that scope or variability is critical to induction. A varied set of premise categories will promote induction, but it is easier to draw an inference about a narrow conclusion category than about a broad conclusion category. Properties in an inductive argument seem to have a breadth or scope of their own, with some properties being restricted to a particular place or time and other properties seeming to generalize easily to many cases. Even 4-year-old children seem to have a sophisticated awareness about the scope of properties (Waxman et al., 1997). The final theme is that similarity is a crucial concept. Just as similarity between premise and conclusion categories promotes induction, dissimilarity within a set of premise

categories also promotes induction. (Indeed, the diversity effect can be thought of as a kind of similarity effect. Because similar categories are expected to share properties, learning that two diverse categories share a property seems more surprising or informative than learning that two similar categories share a property.) Even typicality effects can be explained in terms of similarity, because a category's typicality is highly correlated with its similarity to the representation of its superordinate.

The property effects described here are interesting because they place limits on the use of similarity as an account for inductive inference: Different measures of similarity would be needed for different properties, and in some cases, it is clear that other constructs such as explanations are needed in order to account for human inference. Sloman (1994, 1997) explicitly investigated explanation-based reasoning in induction, and the other studies on property effects (e.g., Heit & Rubinstein, 1994, and Smith et al., 1993) also point, indirectly, to reasoning processes beyond straightforward assessment of similarity.

How do psychological models of induction address these results?

Now that these main results in inductive reasoning have been presented (see Table 1 for a list of key results), it is time to move to the models of induction that have been developed by psychologists. Choosing these models for presentation requires some degree of focus. After all, any theoretical account or explanation of inductive reasoning could be considered a model in some sense. However, for comparability, the focus will be on formal models that are either mathematical or computational descriptions. Also, this section will focus on whether the various models can account for the main results, rather than provide complete presentations of the models themselves (for which the reader is referred to the original sources).

Rips (1975). Chronologically speaking, the first formal model of induction was that of Rips (1975). This modeling effort was performed by deriving multidimensional scaling solutions for different categories of animals, so that similarity and typicality measures could be derived from the animals' positions on a scaling solution. Then Rips applied a set of multiple regression equations to look at various predictors of inductive strength, such as premise–conclusion similarity, premise typicality, and conclusion typicality. The resulting regression model, which included the first two predictors, can account for some of the main results with adult subjects and single-premise arguments—namely, similarity and premise typicality effects.

Potentially, this model could also be applied to some of the developmental trends that have been reviewed, with the assumption that adults and children of different ages would have different multidimensional representations of their knowledge of animals. For example, some of the differences in projection for children and adults reported by Carey (1985) could be explained in terms of *human* being more typical (or central) for children than for adults. Likewise the greater sensitivity to similarity for older children could be captured in terms of greater differentiation in the multidimensional representation for older children, or a greater coefficient for similarity in the regression equation. In principle, this model could be applied to expertise differences as well; the cultural differences found by López et al. (1997), for example, could again be explained in terms of different representations of animal categories being used by American college students and Itzaj subjects.

To evaluate whether this model can account for differences due to development and expertise, it would be necessary to perform multidimensional scaling for the relevant subject population. The model predicts that inductive judgments will be strongly related to these derived similarity measures. However, Medin et al. (1997) did find some dissociations between similarity judgments and inductive judgments for different kinds of tree experts, so the model would have some trouble with these results. Likewise, the results that showed people overriding similarity (e.g., those in Lassaline, 1996; Smith et al., 1993), would be out of bounds for this model.

Without further assumptions, the model does not seem to be sensitive to property effects. For example, Heit and Rubinstein (1994) and Ross and Murphy (1999) showed that different measures of similarity were used for predicting different properties. But if the Rips (1975) model relies on a fixed multidimensional scaling solution, then it would predict the same use of similarity information for different properties. In addition, the model does not really address the difference between projectable and nonprojectable properties, or why permanent characteristics seem to project better than idiosyncratic properties.

Which other results can the Rips (1975) model not address? This model was aimed only at single-premise arguments, so it does not address any of the phenomena with multiple premise categories. Also, the model does not account for one of the most basic results with single-premise arguments—namely, that specificity or homogeneity of the conclusion category promotes induction (Nisbett et al., 1983; Osherson et al., 1990). This model derives its predictions from points represented in multidimensional space. If, for example, *robins* and *birds* are located very near each other in conceptual space because of their similar representations, then the regression model will make similar predictions for inferences about these two categories. In contrast, people will make weaker inferences about the more general category, *birds*. More generally, the model does not make a distinction between categories and individuals. For example, an individual robin would have about the same multidimensional representation as would the *robin* category. Therefore, the model can be applied equally well to reasoning about individuals and about categories, but the model cannot account for any systematic differences that might be found.

Finally, the Rips (1975) model was used to make one of the important discoveries in this area, that typicality of

the conclusion category does not affect inductive strength. However, the model itself does not give an explanation as to why conclusion typicality has no effect. Table 1 shows that the Rips model makes a good start towards addressing Results 1 and 2 and could even account for some group differences such as developmental or expert–novice differences. Otherwise the model does not address these results. Out of fairness, though, it must be said that this model was the first formal psychological account in this area, and it predates most of the results in the table! The Rips model, as will be seen, was influential for subsequent modeling work.

Osherson et al. (1990). The next model of induction, that of Osherson et al. (1990), simultaneously takes a major qualitative leap beyond the Rips (1975) model, now addressing multiple-premise arguments, while at the same time including the Rips model as a special case for single-premise arguments. Just as the Rips model used similarity and typicality as predictors, the Osherson et al. (1990) model has two main components. The first component assesses the similarity between the premise categories and the conclusion category. However, the similarity measure is derived from overlap in a featural representation, rather than from a multidimensional scaling solution. The model can be applied to individuals or to categories, as long they can be described in terms of feature sets. The second component measures how well the premise categories cover the superordinate category that includes all the categories mentioned in an argument. For single-premise arguments, coverage more or less reduces to typicality, but for multiple-premise arguments, coverage gives something closer to a measure of diversity. Coverage is best explained in terms of a series of examples (although Osherson et al. do give a computational formulation):

Squirrels have property X (A)

Cows have property X

Cows have property X (B)

Squirrels have property X

Cows have property X (C)

Tunas have property X

Dogs have property X (D)
Cats have property X

Cows have property X

Dogs have property X (E)
Elephants have property X

Cows have property X

Dogs have property X (F)
Elephants have property X

Roses have property X.

For Arguments A and B, the lowest level superordinate that includes all the categories is *mammal*. Coverage is assessed in terms of the average similarity of the premise category to members of the superordinate. To the extent that cows are more typical mammals than squirrels are, and therefore more similar to other kinds of mammals, Argument B will have greater coverage than Argument A. This is how the model addresses typicality effects. Next, consider Argument C. The lowest level superordinate including all the categories would be *animal* rather than *mammal*. On the average, cows are less similar to various kinds of animals, in comparison with the similarity between cows and just mammals. Therefore, Argument C has worse coverage than Argument B does.

The remaining arguments have multiple premises. In assessing similarity between members of the superordinate category and the multiple premises, only the maximum similarity for any one premise category is considered. So, for Argument D, small mammals tend to be similar to dogs and cats, and large mammals tend not to be similar to dogs and cats. So including *cat* as a premise category does not add much information beyond just having *dog* as a premise category alone. In contrast, for Argument E, some mammals are similar to dogs and other mammals are similar to elephants. Therefore, the *elephant* premise adds information, and the coverage for Argument E is greater than that for Argument D. In this way, Osherson et al.'s (1990) model addresses diversity effects, to the extent that greater coverage is correlated with greater diversity. Finally, the model addresses some exceptions to diversity. For example, in Argument F, the inclusive superordinate category would be *living things* rather than *mammals*. In terms of this much wider category, dogs and elephants do not provide particularly good coverage. Hence there would not be much of a diversity effect for Argument F.

The Osherson et al. (1990) model can address all the single-premise phenomena listed above for the Rips (1975) model, and likewise has many of the same limitations, such as not really addressing property effects at all. But in addition the model can address conclusion specificity to some extent. For example, the model can predict stronger inferences with *bird* as a conclusion category rather than *animal*, to the extent that *animal* suggests a broader superordinate category and a lower measure of coverage for the premise category or categories. With a similar rationale, the model might be applied to the conclusion variability results of Nisbett et al. (1983). For example, a narrow category such as "floridium samples" might be easier to cover than a broader category such as "people in the Barratos tribe." However, further investigation would be needed to see whether the model can address the whole pattern of results. In addition, the model as formulated would not make different predictions for obesity versus skin color of the Barratos. In sum, further assumptions would be needed for this model to fully address the effects of homogeneity of the conclusion category.

Another characteristic of the Osherson et al. (1990) model is that it depends on people generating a useful superordinate to include all the categories presented in an argument. Potentially, different people might generate

different superordinates. Indeed, López et al. (1992) suggested that there could be developmental changes in the ability to generate superordinates, so that children might show more adult-like patterns of reasoning when a superordinate is provided, in comparison with situations where they need to generate their own superordinate. This issue of having to generate a superordinate is also implicit in the Rips (1975) model, where typicality assessments must be made relative to some superordinate category.

The Osherson et al. (1990) model is particularly useful for addressing multiple-premise arguments. The second, coverage-based component is valuable for explaining sample size and diversity effects, and some of the exceptions. It is also appealing to explain any lack of sample size and diversity effects in young children as being due to an underdeveloped mechanism for assessing coverage. More generally, one of the advantages of the Osherson et al. (1990) model over the Rips (1975) model is that it seems to give more of a mechanistic explanation rather than simply provide a means for fitting data. Also, particularly for multiple-premise arguments, the Osherson et al. (1990) model is complex enough and well-specified enough to predict a rich and interesting set of phenomena, profitably addressed by Osherson et al. (1990) themselves.

The Osherson et al. (1990) model gives an account of the first two results in Table 1 and addresses Result 3 to some extent. By assuming that different groups of people, such as children and adults, have different featural representations, the model could account for some group differences in these basic results. The coverage component allows the model to account for sample size and diversity effects, Results 4 and 5. Without further assumptions, the model does not address the remaining results, concerning property effects.

Sloman (1993). This model was implemented as a connectionist network, and perhaps its most important difference from the Osherson et al. (1990) model is that it relies solely on feature overlap without a second mechanism assessing coverage of a superordinate category. Indeed, the Sloman model is especially valuable because it shows how much can be accomplished without this second mechanism, bringing into focus what the second mechanism might actually contribute. The Sloman (1993) model can account for many of the same phenomena as can the Osherson et al. (1990) model, and it likewise has many of the same limitations, so mainly the differences will be covered here. In brief, the way this model works is that premises of an argument are encoded by training the connectionist network to learn associations between input nodes representing the features of the premise categories and an output node for the property to be considered. Then the model is tested by presenting the features of the conclusion category and measuring the activation of the same output node. The model accounts for similarity effects, because training and testing on similar input vectors will lead to strong outputs during testing. The model accounts for diversity effects, because training on a diverse set of categories will tend to strengthen a greater number of connections than will training on a narrow range of categories. It would be interesting to see whether the Sloman model could address the apparent developmental changes in diversity effects that can be accounted for rather naturally by the Osherson et al. (1990) model.

The treatment of typicality effects is somewhat less straightforward. Although Rips (1975) found a distinctive contribution of premise typicality beyond similarity, and, more generally, typicality effects have been one of the most robust findings in inductive reasoning, the Sloman (1993) model does not always predict typicality effects. For arguments with general conclusion categories, for example an argument such as Cows/Mammals being stronger than Squirrels/Mammals, the model would account for any typicality effect in terms of feature overlap. That is, *cows* would be more typical of *mammal* as well as being more similar to the representation of *mammal*, and hence the first argument would be stronger. Although the model does predict some premise–conclusion asymmetries (p. 256), it does not predict an independent effect of premise typicality on arguments with specific conclusion categories (e.g., *dog* rather than *mammal*)—that is, independent of any effect of feature overlap or representation of the conclusion category. More precisely, imagine that category A is more typical than category B, but that these two categories have equal feature overlap to category C. On the basis of the results from Rips (1975), we would expect an argument with the form A/C to be stronger than B/C, but this model would not predict any difference between the two arguments. Indeed, the model seems to predict independent effects of typicality of the conclusion category, a result that has not been reported elsewhere.

One of the advantages of the reliance on feature overlap by the Sloman (1993) model is that it can readily account for nonnormative human results that seem to be heavily influenced by similarity, such as the inclusion fallacy and the inclusion similarity effect. An example of the inclusion similarity effect is that the argument Animals/Mammals seems stronger than Animals/Reptiles, despite the two arguments' being equally valid. The Sloman model accounts for this result readily in terms of greater feature overlap between animals and mammals, whereas the Osherson et al. (1990) model predicts that the two arguments would be equally (and perfectly) strong.

Again, in terms of Table 1, the Sloman (1993) model addresses similarity effects, and to an incomplete extent, typicality effects. Like the Osherson et al. (1990) model, the Sloman model can address some effects of different taxonomic levels of the conclusion category, partly addressing Result 3, but it is not clear whether it fully addresses the effects of conclusion variability as described by Nisbett et al. (1983). The model gives a good account for Results 4 and 5, going beyond the Osherson et al.

Table 2
Sample Application of the Bayesian Model

Hypothesis	Range	Degree of Prior Belief $P(H_i)$	$P(D \mid H_i)$	Posterior Belief $P(H_i \mid D)$
1	**Cow → True** **Sheep → True**	.70	1	.93
2	**Cow → True** Sheep → False	.05	1	.07
3	Cow → False **Sheep → True**	.05	0	.00
4	Cow → False Sheep → False	.20	0	.00

Note—Cases in which the property is true for a category are in boldface.

model in terms of explaining some exceptions to diversity effects. Like the previous two models, without further assumptions the Sloman model does not address property effects, Results 6, 7, and 8.

Smith et al. (1993). Unlike the previous three models which did not really address property effects, the "gap" model of Smith et al. (1993) was explicitly intended to address some of the effects of properties on induction. To illustrate this model, it is best to refer to the example in which the premise that poodles can bite through wire is considered stronger than the premise that German shepherds can bite through wire, for the conclusion that Dobermans can bite through wire. According to the gap model, the first step is that the property *biting through wire* potentiates a set of relevant features or dimensions (e.g., size and strength) and a criterion is set for possessing this property (e.g., a minimum size and strength necessary). Then the premise category is compared with this criterion. In the case of poodles biting through wire, the criterion for the property *biting through wire* would be lowered because there is a large gap between previous beliefs about poodles and what has been expected about biting through wire. The result is that Dobermans biting through wire becomes more plausible, owing to a lowered criterion. In comparison, given the premise about German shepherds, the gap would be so small that beliefs would not change much. This premise would not really lead to changes in the plausibility of the conclusion.

Perhaps what is most appealing about the gap model is that it explicitly includes a stage for potentiating features that are relevant to inferences about a particular property. Unlike in the three previous models, there is no default assumption that different properties will be treated the same. Still, the model does not provide an account of the feature potentiation process, but simply assumes that it would be there. The gap model does include a similarity component as well. Thus the model could account for basic similarity effects, and to an initial extent addresses results such as those of Heit and Rubinstein (1994), who found use of different similarity measures for different properties. However, it is not obvious how the model would capture differences between projectable and nonprojectable properties—that is, why some properties are not projected at all or are just projected randomly. Furthermore, the model does not seem to address typicality effects. Even so, the model does allow for multiple premise categories to be combined and explains sample size effects, but it is unclear whether the model would account for diversity effects or nonmonotonicity effects (see Smith et al., 1993, p. 84).

In sum, in several ways the gap model is an important advance over the Osherson et al. (1990) model, but in other ways the model is somewhat simplified and some key phenomena are left out. In terms of Table 1, the model addresses Results 1, 4, 6, and 8.

Heit (1998). The final model to be discussed is the Bayesian model proposed by Heit (1998). The Bayesian model differs somewhat from the other models in that it perhaps is less of a processing-level account. This model was intended to be a computational-level analysis of what, given certain assumptions, would be normative for inductive inferences. The Bayesian model is an attempt to address normative issues in the spirit of Anderson's (1990) rational analysis of cognition. That is, after specifying the goals of a system, the optimal computational means for attaining these goals are considered. Note that the Bayesian model is by no means an attempt to provide logical justification for inductive inferences or to explain why induction is successful in the real world. It is simply an analysis of the steps that could be taken in a probability estimation task.

According to the Bayesian model, evaluating an inductive argument is conceived of as learning about a property, in particular learning for which categories the property is true or false. For example, in argument

Cows can get disease X

Sheep can get disease X,

the goal is to learn which animals can get this disease and which animals cannot. The model assumes that for a novel property such as the one in this example, people would rely on prior knowledge about familiar properties in order to derive a set of hypotheses about what the novel property might be like. For example, people know some facts that are true of all mammals, including cows and sheep, but they also know some facts that are true just of cows and likewise some facts that are true just of sheep. The question is, Which of these known kinds of properties does the novel property *Can get disease X* resemble most? Is it a cow-and-sheep property, or a cow-only property, or a sheep-only property? To answer this question, the Bayesian model treats the premise or premises in an inductive argument as evidence, which is used to revise beliefs about the prior hypotheses according to Bayes's theorem. Once these beliefs have been revised, the plausibility of the conclusion is estimated.

It will be helpful to present more details of the model in the context of this example. People know quite a few properties of animals, but these known properties must

fall into four types: properties that are true of cows and sheep, properties that are true of cows but not sheep, properties that are true of sheep but not cows, and properties that are not true of either cows or sheep. These four types of known properties can serve as four hypotheses when one is reasoning about novel properties, because any new property must also be one of these four types. These four types of properties are listed in Table 2, with cases in which the property is true for a category shown in boldface for emphasis.

As is shown in Table 2, a person would have prior beliefs about these hypotheses. For example, the value of .70 for Hypothesis 1 represents the belief that there is a 70% chance that a new property would be true of both cows and sheep. This high value could reflect the high degree of similarity between cows and sheep and that people know many other animal properties that are true of both cows and sheep. (The particular numbers are used only for illustration at this point.) However, the person might see a 5% chance that a new property would be true of cows and not sheep, a 5% chance that a new property would be true of sheep and not cows, and a 20% chance that the property is true of neither category. Note that because the four hypotheses are exhaustive and mutually exclusive, their corresponding prior beliefs add up to 1.

This table describes prior beliefs not only about the four hypotheses but also about the two categories. If we combine Hypotheses 1 and 2, it appears that the person believes that there is a 75% chance that cows would have the new property, and likewise if we combine Hypotheses 1 and 3, the person believes that there is a 75% chance that sheep have the new property.

The next step is to combine these prior beliefs with new evidence, using Bayes's theorem. The given premise, *Cows have Property P*, is used to update beliefs about the four hypotheses, so that a better evaluation of the conclusion, *Sheep have Property P*, may be achieved. When we apply Bayes's theorem (Equation 1), the premise is treated as the data, D. The prior degree of belief in each hypothesis is indicated by $P(H_i)$. (Note that there are four hypotheses, so $n = 4$ here.) The task is to estimate $P(H_i \mid D)$—that is, the posterior degree of belief in each hypothesis, given the data.

$$P(H_i \mid D) = \frac{P(H_i)P(D \mid H_i)}{\sum_{j=1}^{n} P(H_j)P(D \mid H_j)} \qquad (1)$$

In Table 2, the calculations are shown for all four hypotheses, given the data that *Cows have Property P*. The calculation of $P(D \mid H_i)$ is quite easy. Under Hypotheses 1 and 2, cows have the property in question, so obtaining the data (that cows have the property) has a probability of 1. But under Hypotheses 3 and 4, cows do not have the property, so the probability of obtaining the data must be 0 under these hypotheses. The final column, indicating the posterior beliefs in the four types of properties, has been calculated with Equation 1. Notably, Hypothesis 1, that cows and sheep have the property, and Hypothesis 2, that just cows have the property, have been strengthened. The two remaining hypotheses have been eliminated from contention, because they are inconsistent with the data or premise that cows have the property.

Finally, the values in Table 2 may be used to evaluate the conclusion, that sheep have Property P. The degree of belief in this conclusion is simply the sum of the posterior beliefs for Hypotheses 1 and 3, or .93. Recall that before the introduction of evidence that cows have the property, the prior belief that sheep have the property was only .75. Thus, the premise that cows have the property led to an increase in the belief that horses have the property.

This illustration raises the important issue of how the prior beliefs, such as the numbers in the third column of Table 2, might be derived. Are the exact values of the priors important? These questions are fundamental issues for Bayesian statistics (see, e.g., Box & Tiao, 1973; Raiffa & Schlaifer, 1961; see also Heit & Bott, 2000). For the purposes of Heit (1998), it was assumed that the priors would be determined by the number of known properties of each type that are brought to mind in the context of evaluating the inductive argument. It might be said that the prior beliefs for new properties are estimated with the use of something like an availability heuristic (Tversky & Kahneman, 1973) based on known properties. The basic idea is that when reasoning about novel animal properties, people would retrieve a set of familiar animal properties from memory. Then they would count up how many known properties are consistent with each of the four properties—for example, how many known properties of animals are true of both cows and horses. The priors in Tables 2, for example, are consistent with the idea that 20 known properties are brought to mind: 14 of Type 1, 1 of Type 2, 1 of Type 3, and 4 of Type 4.

In addition, Heit (1998) argued that the exact values for the prior beliefs are not critical in many cases. For instance, in the present example, the initial degree of belief in Hypothesis Type 4, that neither cows nor horses have the property, was not at all important. The posterior belief in Hypothesis 1, $P(H_1 \mid D)$, can be calculated simply from the prior beliefs in Hypotheses 1 and 2, $P(H_1 \mid D) = P(H_1)/[P(H_1) + P(H_2)]$, or $.93 = .70/(.70 + .05)$. The posterior belief in Hypothesis 1 would be the same regardless of the value of $P(H_4)$, as long as $P(H_1)$ and $P(H_2)$ maintain the same ratio to each other.

The Bayesian model addresses many of the key phenomena reviewed in this paper. For example, the model predicts similarity effects, because novel properties would be assumed to follow the same distributions as would familiar properties. The argument Cows/Sheep seems strong, because many known properties are true of both categories. In contrast, Hedgehogs/Sheep seems weaker, because prior knowledge indicates that there are fewer properties in common for these two categories. The Bayesian model also addresses typicality effects, under the assumption that according to prior beliefs, atypical categories such as *hedgehog* would have a number of

idiosyncratic features. Hence a premise asserting a novel property about hedgehogs would suggest that this property is likewise idiosyncratic and not to be widely projected. In contrast, prior beliefs about typical categories would indicate that they have many features in common with other categories, and hence a novel property of a typical category should generalize well to other categories. (In comparison with the Sloman, 1993, model, the Bayesian model predicts an independent influence of premise typicality, rather than conclusion typicality, beyond feature overlap.)

The Bayesian model also addresses diversity effects, with a rationale similar to that for typicality effects. An argument with two similar premise categories, such as *cows* and *horses*, could bring to mind a lot of idiosyncratic properties that are true just of large farm animals. Therefore a novel property of cows and horses might seem idiosyncratic as well. In contrast, an argument with two diverse premise categories, such as *cows* and *hedgehogs*, could not bring to mind familiar idiosyncratic properties that are true of just these two animals. Instead, the prior hypotheses would be derived from known properties that are true of all mammals or all animals. Hence a novel property of cows and hedgehogs should generalize fairly broadly. This is a quite strong prediction of the Bayesian model, and it is not yet clear how the model would account for any lack of diversity in children. Likewise, it would take further investigation to see whether the Bayesian model would apply to other exceptions to diversity such as the nonmonotonicity effect reported by Osherson et al. (1990) as well as other nonnormative results such as the inclusion fallacy (Osherson et al., 1990) and the inclusion similarity effect (Sloman, 1993, 1998). It could be the case that the Bayesian model has difficulty explaining these nonnormative results. On the other hand, to the extent that people can rely on different priors for answering different questions, the apparent inconsistencies in reasoning might be due to the knowledge that is retrieved for answering particular questions rather than the reasoning process itself.

The Bayesian model can address conclusion homogeneity effects, as in Nisbett et al. (1983). For example, Nisbett et al. found that after a single observation, people were fairly willing to generalize that all floridium samples conduct electricity. The result can be explained in terms of people's initially entertaining two hypotheses: All floridium samples do not conduct electricity, and all floridium samples do conduct electricity. Observing just a single sample of floridium that conducts electricity fits with the second hypothesis and rules out the first hypothesis; hence a strong generalization proceeds rapidly. In contrast, the result for Barratos and obesity was that seeing just one obese Barratos did not promote strong inferences about the whole group. In this case, people might entertain a whole distribution of prior hypotheses—for example, 0% of Barratos are obese, 1% are obese, 2% are obese, . . . , 50% are obese, 51% are obese, . . . , 99% are obese, 100% are obese. Observing one obese Barratos would rule out the 0% hypothesis, and it might cast doubt on the 1% hypothesis; but it would not license the strong inference that all Barratos are obese.

In a similar way, an idiosyncratic property such as *has a scratch on it* could lead people to entertain a diffuse set of prior hypotheses, so that a single observation would not lead to strong inferences. More generally, because the essence of the Bayesian model is that it derives inferences based on prior knowledge of familiar properties, it should be highly sensitive to content effects such as property differences and effects of expertise. The key idea is that the novel property in an argument serves as a cue for retrieving familiar properties. Most psychology experiments on inductive reasoning have used novel properties that sounded at least vaguely biological or internal. In addition, people may retrieve familiar biological properties as a default, for animal categories. So unless the novel property suggests otherwise, people would tend to rely on distributional information about known biological properties.

To give another example, when reasoning about the anatomical and behavioral properties in Heit and Rubinstein (1994), subjects could have drawn on different priors for the two kinds of properties. As in many other experiments, reasoning about anatomical properties led people to rely on prior knowledge about familiar anatomical properties. In contrast, when reasoning about a behavioral property such as *prefers to feed at night*, the prior hypotheses could be drawn from knowledge about familiar behavioral properties. These priors would tend to promote inferences between animals such as hawks and tigers that are similar behaviorally rather than anatomically.

To conclude, the Bayesian model has the potential to address all of the phenomena listed in Table 1. However, the main drawback of this model is that it has not been fully tested. Heit (1998) presented illustrations of how the model might account for a variety of phenomena, but the model has not been directly applied to human data. To test the Bayesian model properly, it would be necessary to collect data about people's beliefs about a large number of familiar properties and then use these data to predict judgments about novel properties. Of course, the other models also depend on collected data such as property listings or similarity ratings, in order to generate predictions. One difference is that the Bayesian model can also respond to beliefs about hidden essences (cf. Medin & Ortony, 1989); for example, the belief that all pieces of limestone have got something unique and distinctive in common, even if one cannot specify exactly what that is. These beliefs might not be easily measured from property listings. Still, the Bayesian model does begin to address a broader range of phenomena than those addressed by the other models.

Discussion. To some extent, there has been a developmental trend among psychological models of induction, with more recent models not surprisingly taking on a wider range of results. Still, perhaps what all the mod-

els have in common is more important than their differences, with some notion of similarity (in terms of feature overlap or proximity in multidimensional space) and some notion of diversity (in terms of category coverage or feature overlap) driving many of the predictions. Given the commonalties among all the models, the main value of the Bayesian analyses by Heit (1998) may be that they highlight the normative basis for the models' predictions. Although it does not address property effects, the Osherson et al. (1990) model has been most influential because it does bring together a lot of phenomena and make interesting predictions of further results.

If one can project from past trends, future models of induction may address content and property effects to a further extent, in light of results showing effects of expertise on induction and widespread property effects even with very young children. Certainly, the wide range of phenomena addressed by the Heit (1998) model, even at an initial stage, should encourage future models to go further. Ideally, future models will give a better process-level account of what Smith et al. (1993) referred to as *feature potentiation*—that is, selecting features that are relevant to a particular inference. Also, to the extent that induction involves explanatory or causal reasoning as suggested by the studies of Sloman (1994) and others, it must be admitted that none of the existing models gives a satisfying account of explanatory reasoning.

Conclusion: Future Directions for Empirical Research

Although much progress has been made in empirical work on inductive reasoning in the past 25 years, by reading between the lines in this review one can see areas of incompleteness that might be profitably investigated in future studies. Perhaps the clearest way to look at inductive reasoning is to do so in terms of the various phenomena, such as typicality effects and diversity effects, that appear in Table 1. As one considers these results, it is natural to be interested in whether they appear in different groups of people, such as children or adults, and Western cultures or non-Western or traditional cultures. Work that addresses such questions is well under way, although there are still many interesting questions to be addressed, such as why diversity-based reasoning seems to be harder to find in some groups.

Another way to think about induction is to do so in terms of the various tasks and responses that would require inductive reasoning. For example, in the experiments described in this review researchers have used response measures such as probability judgments, judgments of inductive strength, forced-choice predictions, and behaviors such as how an infant plays with a toy. The tasks varied in another important way as well. In some experiments, the premises gave information about individuals (e.g., a particular bird has some property) and in other experiments, the premises gave information about categories (e.g., a kind of bird has some property). Possibly there was even some ambiguity in some experiments whether the premises referred to individuals or categories. This problem could particularly come up when premises are presented in picture form, if it is unclear whether a picture of some individual is meant to stand for a class of items. It would be important to establish whether the various phenomena of inductive reasoning, listed in Table 1, do appear for different versions of the task. For example, all of the models of induction described here can apparently be applied to inferences about individuals or categories. Systematic research could potentially show differences in reasoning about individuals as opposed to categories, and these differences might or might not correspond to the models' predictions.

A related issue is how well the laboratory-based tasks reported here match up to inductive reasoning as manifested by everyday judgments and decisions. How well do the phenomena reviewed here correspond to everyday reasoning? It is hoped that most of the key results in Table 1 would occur outside of the laboratory as well. Perhaps the more contentious results would be the fallacies reported by Osherson et al. (1990) and Sloman (1993, 1998), in which people violate basic laws of probability. It would be valuable to study whether these reasoning fallacies are robust enough to appear in the real world and in everyday choices, or whether they are dependent on the characteristics of experimental settings and survey methodology.

Still another way to think about the phenomena of inductive reasoning is to consider whether they might be different for various domains of knowledge or different kinds of categories—for example, natural kinds, artifacts, social categories, event categories, ad hoc categories (Barsalou, 1983). Again, this is an intriguing possibility that would need to be investigated more systematically in future research. Some studies have been done with different kinds of categories, but the majority of published experiments have used animal categories (and animals' biological properties). Although there could be many reasons for this focus on categories of animals, the risk remains that the results might be different in other domains or with other kinds of categories. It is unclear whether the emphasis on animal categories in published papers simply reflects the choices of experimenters in creating stimuli, or whether there is some nonpublication bias because experiments with other stimuli did not yield interpretable results.

Therefore, after 25 years of psychological research on inductive reasoning, it is time both to acknowledge the extensive progress that has been made, especially in terms of the regularities that have been documented, and to acknowledge that future empirical work needs to be more ambitious, ideally guided by more ambitious models as well.

REFERENCES

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

BACON, F. (1898). *Novum organum*. London: George Bell and Sons. (Original work published 1620)

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

BARSALOU, L. W. (1983). Ad hoc categories. *Memory & Cognition*, **11**, 211-227.

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

BOX, G. E. P., & TIAO, G. C. (1973). *Bayesian inference in statistical analysis*. London: Addison-Wesley.

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

CHOI, I., NISBETT, R. E., & SMITH, E. E. (1997). Culture, category salience, and inductive reasoning. *Cognition*, **65**, 15-32.

COLEY, J. D., MEDIN, D. L., & ATRAN, S. (1997). 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. B., & ATRAN, S. (1999). Inductive reasoning in folkbiological thought. In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 205-232). Cambridge, MA: MIT Press.

FLORIAN, J. E. (1994). Stripes do not a zebra make, or do they: Conceptual and perceptual information in inductive inference. *Developmental Psychology*, **30**, 88-101.

FRIED, L. S., & HOLYOAK, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 234-257.

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: Categories and inferences in 2-year-old children. *Developmental Psychology*, **26**, 796-804.

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.

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

GOODMAN, N. (1972). *Problems and projects*. Indianapolis: Bobbs-Merrill.

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

HADJICHRISTIDIS, D., SLOMAN, S. A., STEVENSON, R. J., & OVER, D. E. (1999). Centrality and property induction. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society* (p. 795). Mahwah, NJ: Erlbaum.

HAHN, U., & CHATER, N. (1997). Concepts and similarity. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 43-92). London: Psychology Press.

HEIT, E. (1997). Knowledge and concept learning. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 7-41). London: Psychology Press.

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: Oxford University Press.

HEIT, E., & BOTT, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 39, pp. 163-199). San Diego: Academic Press.

HEIT, E., & HAHN, U. (1999). Diversity-based reasoning in children age 5 to 8. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society* (pp. 212-217). Mahwah, NJ: Erlbaum.

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

HEMPEL, C. G. (1966). *Philosophy of natural science*. Englewood Cliffs, NJ: Prentice Hall.

HOMA, D., & VOSBURGH, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 322-330.

HUME, D. (1988). *An enquiry concerning human understanding*. La Salle, IL: Open Court. (Original work published 1748)

KALISH, C. W., & GELMAN, S. A. (1992). On wooden pillows: Multiple classifications and children's category-based inductions. *Child Development*, **63**, 1536-1557.

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

LOOSE, J. J., & MARESCHAL, D. (1999). Inductive reasoning revisited: Children's reliance on category labels and appearances. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society* (pp. 320-325). Mahwah, NJ: Erlbaum.

LÓPEZ, A. (1995). The diversity principle in the testing of arguments. *Memory & Cognition*, **23**, 374-382.

LÓPEZ, 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.

LÓPEZ, A., GELMAN, S. A., GUTHEIL, G., & SMITH, E. E. (1992). The development of category-based induction. *Child Development*, **63**, 1070-1090.

MACARIO, J. F., SHIPLEY, E. F., & BILLMAN, D. O. (1990). Induction from a single instance: Formation of a novel category. *Journal of Experimental Child Psychology*, **50**, 179-199.

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

MANDLER, J. M., & MCDONOUGH, L. (1998). Studies in inductive inference in infancy. *Cognitive Psychology*, **37**, 60-96.

MANDLER, J. M., & MCDONOUGH, L. (in press). Advancing downward to the basic level. *Journal of Cognition & Development*.

MCDONALD, J., SAMUELS, M., & RISPOLI, J. (1996). A hypothesis-assessment model of categorical argument strength. *Cognition*, **59**, 199-217.

MEDIN, D. L., GOLDSTONE, R. L., & GENTNER, D. (1993). Respects for similarity. *Psychological Review*, **100**, 254-278.

MEDIN, D. L., LYNCH, E. B., COLEY, J. D., & ATRAN, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *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: Cambridge University Press.

MILL, J. S. (1874). *A system of logic*. New York: Harper.

MURPHY, G. L., & ROSS, B. H. (1999). Induction with cross-classified categories. *Memory & Cognition*, **27**, 1024-1041.

NAGEL, E. (1939). *Principles of the theory of probability*. Chicago: University of Chicago Press.

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.

OSHERSON, D. [N.], SMITH, E. E., MYERS, T. S., SHAFIR, E., & STOB, M. (1994). Extrapolating human probability judgment. *Theory & Decision*, **36**, 103-129.

OSHERSON, D. N., SMITH, E. E., WILKIE, O., LÓPEZ, A., & SHAFIR, E. (1990). Category-based induction. *Psychological Review*, **97**, 185-200.

OSHERSON, D. N., STERN, J., WILKIE, O., STOB, M., & SMITH, E. E. (1991). Default probability. *Cognitive Science*, **15**, 251-269.

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. L., & MEDIN, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 811-828.

RAIFFA, H., & SCHLAIFER, R. (1961). *Applied statistical decision theory*. Boston: Harvard University, Graduate School of Business Administration.

RIPS, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning & Verbal Behavior*, **14**, 665-681.

ROSCH, E., MERVIS, C. G., GRAY, W. D., JOHNSON, D. M., & BOYES-

BRAEM, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382-439.

ROSS, B. H., & MURPHY, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, **38**, 495-553.

SHIPLEY, E. F. (1993). Categories, hierarchies, and induction. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 30, pp. 265-301). San Diego: Academic Press.

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*, **2**, 81-110.

SLOMAN, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, **35**, 1-33.

SMITH, E. E., SHAFIR, E., & OSHERSON, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition*, **49**, 67-96.

SPELLMAN, B. A., LÓPEZ, A., & SMITH, E. E. (1999). Hypothesis testing: Strategy selection for generalising versus limiting hypotheses. *Thinking & Reasoning*, **5**, 67-91.

SPRINGER, K. (1992). Children's awareness of the biological implications of kinship. *Child Development*, **63**, 950-959.

TVERSKY, A., & KAHNEMAN, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, **5**, 207-232.

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.

WU, M., & GENTNER, D. (1998). Structure in category-based induction. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1154-1158). Mahwah, NJ: Erlbaum.

NOTE

1. Quite a few studies, including parts of Carey (1985), have looked at attribution tasks rather than projection tasks. In an attribution task, the subject, typically a child, states whether some familiar item has some familiar property, such as whether dogs sleep. The contribution of inductive reasoning to attribution tasks is unclear, because in many cases the subject would be able to answer on the basis of established knowledge or observations without a major role for inductive inference. Therefore, this paper will focus on projection tasks, which involve unfamiliar categories and/or properties.

(Manuscript received March 18, 1999;
revision accepted for publication February 22, 2000.)