Cognitive Psychology **43,** 243–273 (2001)
doi:10.1006/cogp.2001.0757, available online at http://www.idealibrary.com on IDEAL®

Diversity-Based Reasoning in Children

Evan Heit

University of Warwick

and

Ulrike Hahn

Cardiff University

One of the hallmarks of inductive reasoning by adults is the diversity effect, namely that people draw stronger inferences from a diverse set of evidence than from a more homogenous set of evidence. However, past developmental work has not found consistent diversity effects with children age 9 and younger. We report robust sensitivity to diversity in children as young as 5, using everyday stimuli such as pictures of objects with people. Experiment 1 showed the basic diversity effect in 5- to 9-year-olds. Experiment 2 showed that, like adults, children restrict their use of diversity information when making inferences about remote categories. Experiment 3 used other stimulus sets to overcome an alternate explanation in terms of sample size rather than diversity effects. Finally, Experiment 4 showed that children more readily draw on diversity when reasoning about objects and their relations with people than when reasoning about objects' internal, hidden properties, thus partially explaining the negative findings of previous work. Relations to cross-cultural work and models of induction are discussed.

Key Words: inductive reasoning; children; diversity; evidence.

One of the most important functions of categories is that they allow us to make predictions and draw inferences. For example, in seminal work by Rips (1975), subjects drew inferences from one category of animals to another. They were told to imagine an island where all members of one category of

We are grateful to Jane Pollock, Emma Bourne, and Georgia Barnett for assistance in conducting this research. We thank the students and teachers of St. Peter's Primary School, Brookhurst Primary School, Cubbington Primary School, and Cannon Park Primary School for their participation. John Coley, Susan Gelman, Brett Hayes, Denis Mareschal, Douglas Medin, and Daniel Osherson provided valuable comments on this work. A preliminary version of Experiments 1 and 2 appeared in Heit and Hahn (1999), and this work was also presented at the London meeting of the Experimental Psychology Society, January 2000.

Address correspondence and reprint requests to Evan Heit, Department of Psychology, University of Warwick, Coventry CV4 7AL, United Kingdom. E-mail: E.Heit@warwick.ac.uk.

0010-0285/01 $35.00

animals, such as rabbits, have a particular disease, then they estimated the proportion of another animal category, such as dogs, that would also have this disease. Rips found a predominant tendency toward similarity-based reasoning, namely people were highly sensitive to the similarity between the given (also referred to as premise) category and the target (also referred to as conclusion) category. As an example, subjects made stronger inferences from rabbits to dogs than from rabbits to bears. Consistent with proposals from philosophy (e.g., Mill, 1874), it seems that similarity between a premise category and a conclusion category is a crucial determinant of the strength of an inductive inference. Also, Rips showed typicality effects. The typicality of the premise, with respect to its superordinate category, was critical in promoting inferences. More typical premise categories led to stronger inferences than atypical premise categories. For example, with the bird stimuli, having bluejay as a premise category led to stronger inferences overall compared to having goose as a premise category.

A limitation of this early work is that it only looked at inferences from a single premise. In contrast, people will often face multiple sources of evidence or multiple categories when drawing an inference. Experimental research on inductive inference from multiple categories could be especially revealing about the underlying processes because the richer data set could be particularly constraining on possible theoretical accounts (see Heit, 2000, for a review). The most extensive and influential work on induction from multiple categories was conducted by Osherson et al. (1990). They reported several phenomena involving reasoning with multiple premise categories, but we focus on what is perhaps the most basic phenomenon, which we refer to as the diversity effect. This phenomenon is illustrated by the following example. In this notation, the statements above the line are premises, which are assumed to be true, and the task is to assess the strength of the conclusion statement, which is below the line.

(1) Lions have an ulnar artery.
Giraffes have an ulnar artery.
Rabbits have an ulnar artery.

(2) Lions have an ulnar artery.
Tigers have an ulnar artery.
Rabbits have an ulnar artery.

People tend to find arguments like (1) to be stronger than arguments like (2), even though giraffes are very different from rabbits. What is critical is the diversity of the premise categories. For argument (1), lions and giraffes are such a diverse set of premise categories that it seems to license a broad set of inferences, such as that rabbits and many other mammals have an ulnar artery as well. In contrast, for argument (2), lions and tigers are a very nondiverse set, and it seems possible that the property of interest, having an ulnar artery, could be restricted to just these two animals or just to felines.

Supported by these diversity effects in adults, Osherson et al. (1990) developed a computational model of induction that includes not only a similarity-based component but also a coverage-based component, in which people generate an inclusive category and assess how well the premise categories cover this superordinate. In the present example, lions and giraffes would cover the inclusive category, mammals, better than lions and tigers, hence a stronger inference to other mammals would be indicated. This account by Osherson et al. not only describes a fairly sophisticated reasoning procedure but also presupposes knowledge of the relevant taxonomic category structure.

Because the diversity effect seems to be highly revealing both about reasoning mechanisms as well as categorical knowledge, there has been keen interest among researchers in assessing the generality of this phenomenon. How robust is diversity-based reasoning? Lopez (1995) devised a more demanding test of diversity-based reasoning, in which people choose premise categories rather than simply evaluate arguments given a set of premises. In other words, will people's choices of premises reveal that they value diverse evidence? Subjects (American college students, as in Osherson et al., 1990) were given a fact about one mammal category and were asked to evaluate whether all mammals have this property. In aid of this task, subjects were allowed to test one other category of mammals. For example, subjects would be told that lions have some property, 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 Lopez (1995) that for inductive arguments about animals, subjects do make robust use of diversity in not only evaluating evidence but also in seeking evidence. The prevalence of diversity effects is less clear for other subject populations, however. In particular, developmental work has generally failed to find diversity effects in children. (See the General Discussion for a comparison to cross-cultural research.)

Past Developmental Results

In fact, the first study of diversity-based reasoning was a developmental one by Carey (1985), which compared 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. In this study, subjects were asked whether various specific kinds of animals, presented in pictures, would also have this property. The purpose of this study was to see whether subjects reason that ''if two such disparate animals as dogs and bees'' have some 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 infer-

ences 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.

In a follow-up experiment, Carey (1985) looked at diversity effects based on the concept of living thing rather than animal. The results were somewhat less clear for this study. Children taught a fact about dogs and flowers tended to generalize more broadly than children taught a fact about dogs and bees, but the pattern of generalization was overextended and not adultlike. In particular, quite a few children also attributed the fact to inanimate objects. Again, there was no definitive evidence for mature diversity-based reasoning in children. Carey's interpretation was mainly in terms of knowledge differences, with children having a less well-developed living thing concept than adults. However, looking at both experiments together, the results also point to processing differences between children and adults. For example, in terms of the Osherson et al. (1990) model, children's mechanism for generating an inclusive category and assessing coverage does not appear to be fully developed.

The experiments by Lopez, Gelman, Gutheil, and Smith (1992) were an attempt to translate the written arguments used in adult studies by Osherson et al. (1990) to children's versions with pictures. For example, to look at similarity effects, children were shown a picture of a horse and told that it has leukocytes (or some other unfamiliar property) inside. Then they were shown pictures of a donkey and squirrel and asked which one would also have leukocytes. To look at typicality effects, children were told that dogs have leukocytes inside and bats have ulnaries inside. Then they were asked whether ''all animals'' have leukocytes inside or ulnaries inside. With this procedure, children as young as 5 years showed robust similarity and typicality effects. However, the evidence for diversity effects was rather limited.

To investigate diversity effects, children were given a diverse set of animals (cats and buffaloes) and a nondiverse set (cows and buffaloes). The children were given one property for each set, e.g., that cats and buffaloes have leukocytes inside, and cows and buffaloes have ulnaries inside. When asked whether another specific animal, such as kangaroo, would have leukocytes or ulnaries inside, 5-year-olds and 9-year-olds showed chance performance. They were not influenced by the greater diversity of the set of animals with leukocytes. The results were somewhat different when children were asked whether ''all animals'' have leukocytes or ulnaries inside. Again, 5-year-olds were at a chance level of performance. Even with a modified procedure in which the experimenter emphasized the diversity of one set and the nondiversity of the other, 5-year-olds still did not show diversity effects. However, 9-year-olds did show a diversity effect with the general conclusion category ''all animals.''

The results for 5-year-olds clearly did not show diversity effects, and the mixed results for 9-year-olds were interpreted by Lopez et al. (1992) in terms

of the Osherson et al. (1990) model. In particular, Lopez et al. concluded that the mechanism for generating inclusive categories was not fully developed in 9-year-olds. Therefore, they could show limited diversity effects when the covering category (''all animals'') is given, but they do not show diversity effects when drawing a conclusion about any specific mammal.

Gutheil and Gelman (1997) made a further attempt to find evidence of diversity-based reasoning for specific conclusion categories in 9-year-olds. Their basic design corresponded to that of Lopez et al. (1992): There was always a diverse set of items (e.g., five different butterflies) and a nondiverse set of items (e.g., five similar butterflies). Children were asked about another item such as another butterfly. However, category members at lower, or more concrete, taxonomic levels were used which would presumably make reasoning easier (see, e.g., Gelman, 1988). In this study, which was also intended to look at sample size effects, sample sizes were increased (e.g., five butterflies rather than two mammals). In addition, Gutheil and Gelman used pretested properties that were expected to promote systematic reasoning; adult subjects had shown sample size effects in reasoning about these properties. These properties were simpler and more concrete, e.g., ''has a red spot under its wing'' rather than ''has leukocytes inside.'' Still, like Lopez et al. (1992), Gutheil and Gelman did not find diversity effects for specific conclusion categories with 9-year-olds, although in a control condition with adults, there was clear evidence for diversity effects using the same stimuli. Gutheil and Gelman's assessment of 9-year-olds' inductive abilities was more negative than that of Lopez et al. Gutheil and Gelman assumed that when the stimuli all come from the same basic-level category (e.g., butterflies or birds), 9-year-old children should be able to recognize the inclusive category. Hence, the failure of 9-year-olds to show diversity effects was interpreted not in terms of a difficulty in generating an inclusive category, but rather ''a more entrenched difficulty using . . . diversity information in their inductive judgments'' (p. 172).

A recent study by Rozelle, Sides, and Osherson (1999) also gave a negative picture of children's ability to use diversity in induction. Children, ages 9 to 11 years, were given arguments with general conclusion categories. For example, they judged which gave stronger evidence that all mammals have ascorbic acid in their bodies, that rabbits and squirrels have ascorbic acid, or that rabbits and cows have ascorbic acid. The children were at a chance level in choosing between the nondiverse set and the diverse set. Using a general conclusion category, which seemed to support diversity-based reasoning in Lopez et al. (1992), and with older children than the other developmental studies, there was still no evidence for diversity-based reasoning.

Rationale

The negative findings with respect to diversity effects in children from 5 to 9 years old, and older, are surprising for a number of reasons. This period is critical for school learning and real-world observation of many categories

such as different kinds of plants and animals. The diversity effect is one of the more basic phenomena of inductive reasoning with more than one piece of evidence. (Osherson et al., 1990, detailed a number of more subtle phenomena.) Do 9-year-old children process multiple pieces of evidence in a qualitatively different manner than adults? As we review under General Discussion, there is a long history of philosophical and statistical arguments for why it seems normative to draw stronger or broader inferences from diverse evidence than from nondiverse evidence. Are 9-year-old children incapable of using diversity informative for inductive reasoning in a normative manner?

The influence of diversity on inductive reasoning also seems closely related to the influence of variability on categorization. The key finding with adults, that more variable observations promote broader or stronger generalizations, has been well established (e.g., Fried & Holyoak, 1984; Homa & Vosburgh, 1976; Posner & Keele, 1968). There has also been evidence that infants are sensitive to variability of categories (e.g., Mareschal, French, & Quinn, 2000; Quinn, Eimas, & Rosencrantz, 1993; Younger, 1985) in less demanding tasks such as simply looking at stimuli and having looking time measured. Although there are likely many differences between habituation tasks and explicit inductive reasoning tasks, these results seem to rule out the possibility that children as old as age 9 have trouble in perceiving variability or diversity. The main question we address in this article is whether children can incorporate this information into inductive reasoning.

There are several possible explanations for why children in past studies did not show consistent diversity effects. We see at least three interesting ways of explaining the lack of diversity effects in children. First, there could be a developmental change in the mechanisms of inductive reasoning; this explanation was elaborated by Lopez et al. (1992) and Gutheil and Gelman (1997). Reasoning in children might not be able to access all the same processes as reasoning by adults. The inductive reasoning processes used by children just may not be sensitive to diversity information at all, regardless of other details of the task.

Second, there could be a change in knowledge structures; this explanation was the focus of Carey (1985). It could be the case that children do not have fully developed concepts of animals and the taxonomic structure that relates various animals to each other. Hence it would be difficult to be sensitive to the diversity of a set of animal categories with respect to a superordinate. This explanation raises the possibility that performance might be different for other content domains.

A third class of explanations refers to other performance issues. If there are some other difficulties with the experimental tasks, then they may not fully reveal children's abilities. Because our own experiments were modeled on the procedure of Lopez et al. (1992) and Gutheil and Gelman (1997), we focus here on these studies. Although theirs was a natural way to convert

the stimuli for adult studies to children's experiments, this procedure possibly led to other difficulties. For example, children might have difficulties in reasoning about unfamiliar predicates such as ''has leukocytes inside.'' However, Gutheil and Gelman used somewhat more familiar predicates compared to those of Lopez et al. Also, children might have difficulties reasoning about hidden or invisible things such as the hidden properties in both studies. Surely, the reason for using pictures in the children's experiments was to make the task easier or more understandable. By this same logic, making the properties hidden rather than showing them in pictures could have made the task more difficult for children. Finally, these studies generally gave children an intermixed set of questions, such as a similarity question followed by a diversity question. The best strategy for answering one question might be the wrong strategy for the next question. Indeed, the only way that Gutheil and Gelman were able to find a slight amount of evidence for diversity effects in 9-year-olds was to prevent similarity-based strategies. Despite these possible difficulties, we should point out that both papers do present a fairly systematic picture of inductive reasoning. Lopez et al. in particular used these procedures to show good performance by 5-year-olds on similarity and typicality questions. So these methods clearly are valuable, but it is still possible that by making the procedure somewhat easier, any less robust phenomena such as diversity might have a better chance to be evidenced.

Our own experiments were an attempt to distinguish between the first explanation, processing differences, and the second and third explanations, content influences, and performance issues. The main aim of our experiments was to show that at some level, children can indeed respond to diversity information in an inductive reasoning task in a fairly adultlike manner. To this end, we sought to make the task as simple as possible. Therefore we used materials with a different content than the past studies, everyday objects rather than biological categories of living things, which we expected might lead to different results. Consequently our experiments do not directly address biological reasoning or knowledge as did the past studies. In addition, we simplified the task in a number of practical ways. The test questions were given in blocks all of the same type, instead of intermixing different types of questions. Rather than using hidden properties, for the most part we used visible relations between an object and a person. For example, rather than a cow with leukocytes inside, there might be a doll belonging to Jane, where both the doll and Jane are shown in photographs. To address the general issue of whether children can use diversity information in inductive reasoning, it was not necessary to use hidden properties. Instead we looked at reasoning about a highly familiar relation, possession, on the assumption that children might be at their best when reasoning about people's belongings.

In other ways our procedure was analogous to that used by Lopez et al. (1992) and Gutheil and Gelman (1997). The child was given a fact about a diverse set of items and another fact about a nondiverse set of items. Then

the child was asked which fact was true of a target item. In our first experiment, the facts were always related to possession or other interaction with humans. For example, children were shown a set of three different dolls and told that these dolls belong to a girl named Jane. This was a diverse set of dolls, including a china doll, a stuffed doll, and a Cabbage Patch doll. The set was presented as three pictures of Jane playing with the dolls. Then the children were shown another set of dolls, all the same (three pictures of Barbie dolls). The child was shown that these dolls all belong to a girl named Danielle. Then the child was shown a target item, a baby doll. The question was whether this doll belonged to Jane (the diverse choice) or Danielle (the nondiverse choice). Our stimulus design was analogous to the past studies of diversity in terms of using a diverse set and a nondiverse set and giving one fact for each set. We tested children over a range of ages from 5 to 9 years, with the aim of looking for some evidence of diversity-based reasoning at age 9 and below. In general, we used similar instructions to Lopez et al. and Gutheil and Gelman.

Our experiments focused on inductive inferences to specific items, e.g., inferences to another doll, because the evidence for diversity with specific conclusions up until now has been the scarcest. Gutheil and Gelman (1997) and Carey (1985) took a similar approach, focusing on inductive inferences to specific items rather than general categories (which could not be captured easily by picture stimuli).

The first experiment was simply an attempt to look for some evidence of diversity effects with everyday objects such as dolls. Also, in this experiment, we used two types of instructions, following Lopez et al. (1992). For the first four items in each session, the child was given standard instructions that did not refer to diversity. Then for the last four items, the experimenter was emphatic in noting that one set was diverse and the other was not.

EXPERIMENT 1

Method

Subjects. There were 64 children: 18 in year 1 (mean age 5;7, range 5;3 to 6;0), 19 in year 2 (mean 6;9, range 6;3 to 7;2), 13 in year 3 (mean 7;9, range 7;3 to 8;1), and 14 in year 4 (mean 8;7, range 7;8 to 9;1). All attended St. Peter's Primary School in Leamington Spa, a town in the midlands of England. The experiment was conducted on individual students; each session typically lasted 10–15 min.

Materials. There were eight test questions. For each question, there were two sets of given items as well as a target item, all presented as individual photographs. The given information consisted of a set of three nondiverse items and a set of three diverse items. Each set was associated with a person. For example, in a nondiverse set there were three photographs of a football (soccer ball) being played with by a boy named Tim. In the corresponding diverse set, there were three photographs of a basketball, a cricket ball, and a tennis ball, each being played with by another boy, named Robby. The target item was a picture of another item from the same general category, such as a photograph of a rugby ball. This photograph was

TABLE 1
Stimuli for Experiment 1

Target item	Nondiverse set	Diverse set
Rugby ball	Football (soccer ball)	Basketball, cricket ball, tennis ball
Baby doll	Barbie doll	China doll, stuffed doll, Cabbage Patch doll
Purple brimmed hat	White floppy hat	Straw hat, ski hat, baseball hat
Red top (shirt)	Green top	Blue top, white top, gray top
White book	Red book	Black book, green book, purple book
Yellow flower	Purple flower	Orange flower, blue flower, red flower
Blackcurrant ice cream (cone)	Chocolate ice cream	Strawberry ice cream, vanilla ice cream, pistachio ice cream
Horse	Cat	Dog, guinea pig, goldfish

Note. These descriptions of objects are summaries. Different objects tended to differ in size, shape, and coloring.

of the item alone, without any person. The test question was to choose whether the target item would go with one person or the other, e.g., Tim or Robby.

The color photographs were mounted on cards approximately 15 × 20 cm. The photographs for each test question used a different pair of people. The stimuli are described briefly in Table 1. We tried to choose diverse sets of items that would be as variable as possible, along multiple dimensions, while remaining within the same category. For example, the diverse set of hats varied in terms of color, size, and shape. Likewise, the target item was chosen to be as different as possible from the items in the nondiverse set and the diverse set in terms of color, size, and shape. Therefore it was not expected that subjects would draw inferences on the basis of simple perceptual similarities between pairs of items.

Procedure. The order of test questions was randomized for each subject, and likewise on half the questions the nondiverse pictures were presented before the diverse pictures and half the time presentation was in the opposite order.

Four test questions were given with standard instructions, and then four test questions were given with emphatic instructions. The standard instructions involved presenting the three nondiverse photographs and three diverse photographs, briefly describing each picture. For example, the experimenter would say, ''Look, there's my friend Tim. He's playing with a football.'' The emphatic version of the instructions increased the salience of nondiversity or diversity. For example, the experimenter would say ''Look, there's Tim. He's playing with the same thing, another football'' for the nondiverse set. Likewise, for the diverse set the experimenter would emphasize the differences between items in the diverse set. The purpose of this manipulation was that for the last four test questions, we wanted to be certain that the diversity or nondiversity of each set was highly salient for each subject.

After the six given pictures were presented, the child was shown the target photograph and asked who this item would go with, e.g., who would play with this item or who would wear it. The experimenter provided mildly positive nondirective feedback after the child's response. Otherwise, the experimenter never gave any reason for the child to favor either the nondiverse set or the diverse set in making inferences in the standard instructions as well as the emphatic instructions. At the end of the eighth question, the experimenter asked the child to explain his or her response.

Results and Discussion

Overall, as shown in Table 2, children robustly favored the diverse choice over the nondiverse choice. For example, in the standard condition, the over-

TABLE 2
Proportion of Diverse Choices, Experiment 1

Year	Standard (*SE*)	Emphatic (*SE*)
1	.71	.78
	(.05)	(.08)
2	.64	.86
	(.09)	(.06)
3	.73	.88
	(.06)	(.05)
4	.93	.91
	(.06)	(.04)

all proportion of diverse choices was .74, which was significantly greater than a chance level of 50%, $t(63) = 6.65$, $p < .001$. Inspection of Table 2 suggests that the emphatic condition led to an even higher level of diverse choice and that there was a tendency for older children to make more diverse choices. A two-way analysis of variance (ANOVA) indicated a significant effect of instructions, $F(1, 60) = 8.66$, $MSE = .05$, $p < .01$. The effect of year was not quite statistically significant, $F(3, 60) = 2.19$, $MSE = .20$, $p < .10$, and the interaction was not close to the level of significance, $F(3, 60) = 1.31$, $MSE = .05$. The age-related trend had some further support from a finer grained analysis, which correlated each child's age with his or her overall proportion of diverse choices. This correlation was .22, $p < .05$, suggesting that older children did indeed make more diverse choices.

The main result, showing diversity effects overall, was consistent across the eight sets of items, with mean proportion of diverse choices ranging from .69 to .90. Therefore, the basic effect does not seem to depend on accidental characteristics of an individual set of items, such as inadvertent similarities between pictures. For none of the eight sets of items did the results go in the opposite direction of diversity effects. This consistency across different kinds of stimuli, from toys to clothing to foods, contrasts with the past results showing a lack of diversity effects for biological properties of living things. Using social properties (human interaction or possession) rather than biological properties, we found diversity effects for living things such as flowers and pet animals.

The children's explanations of their last responses pointed clearly to diversity-based reasoning. We applied a strict criterion to the coding these explanations. An explanation was only considered to be diversity-based if (1) the child actually chose the diverse set on the last test question and (2) the explanation explicitly mentioned diversity or nondiversity. Examples of explanations referring to diversity are ''She's always wearing different hats,'' ''She likes all kinds of tops,'' and ''Joe always eats the same flavor.'' Of 64 children, 50 gave diversity-based explanations. For one child, no explanation

was recorded. The remaining 13 children gave other explanations. Some of these explanations were unclassifiable, but others were similarity-based, i.e., they referred to some element of similarity between the target item and one or more given items. Examples of these other explanations include ''She likes lighter colors'' and ''Rugby is a bit like football (soccer).'' Of the 13 children who gave other explanations, 11 chose the nondiverse set rather than the diverse set. Therefore, the explanations suggest that there was some degree of similarity-based reasoning, but that this similarity-based reasoning did not contribute to the main finding in favor of diverse choices. Instead, similarity-based explanations were associated with choosing the nondiverse set. Overall, the pattern of explanations supports the idea that children were truly responding to diversity and that any inadvertent similarities between pictures tended to work against the overall diversity effect.

These results represent the first strong evidence for diversity-based reasoning in children up to age 9, for specific conclusion items. Indeed, we did not find major age differences in the range of 5 to 9 years—even the 5-year-olds showed diversity effects. On the first four test questions, which did not emphasize the diversity or nondiversity of given items, children overall made the diverse choice 74% of the time. The proportion was even higher with emphatic instructions that highlighted diversity and nondiversity (but did not indicate which one to choose). The fact that highlighting diversity just made the results stronger again suggests that we truly did observe an effect due to diversity rather than to some other accidental properties of all the stimuli. The apparent effect of instructions, however, could also be related to practice because the emphatic instructions were always for the last four items. (We could not present the emphatic instructions for the first four items because these instructions could carry over to affect subsequent performance on later items.)

EXPERIMENT 2

Given the result of Experiment 1, that children as young as age 5 do show diversity effects, we next set out to determine whether their use of diversity is appropriately constrained. Having a diverse set of premise categories should license a broad set of inferences compared to a nondiverse set of premises, but it does not license just any inference at all. Do children have a sense of the reasonable scope of inferences, or in Experiment 1 were they simply choosing the more diverse set without a full understanding of the nature of the task? We attempted to address this issue by choosing target stimuli that would not necessarily license strong inferences from a diverse set. In particular, diverse premise categories should have less of an effect on remote conclusion categories, matching the premise categories only at the superordinate level. For example, again the diverse set of dolls belonged to Jane, and the nondiverse set of dolls belonged to Danielle. But sometimes the subjects

were asked about a yo-yo rather than another doll (a baby doll). The yo-yo matched the premise categories at a more superordinate level than did the doll, which was a basic-level match. If children have a sophisticated sense of diversity and the scope of inferences, then the diversity effect should be attenuated or even eliminated for the more superordinate target items. For adult intuitions at least, just because Jane plays with a wide range of dolls, that does not make her more likely to play than Danielle with a yo-yo.

Furthermore, this prediction is made by the model of Osherson et al. (1990), because a conclusion item that matches the premise items at a superordinate level would lead the subject to generate a very broad category, such as all toys, for the basis of assessing diversity. Neither set of premise categories, even three different dolls, would seem particularly diverse in terms of the space of all toys. Hence the difference in diversity for the two sets of premise categories would be very minor and less likely to affect choices.

We ran a version of this experiment on 12 adults, comparing responses for four basic-level matches (e.g., another kind of doll) and four superordinate-level matches (e.g., another kind of toy). The adults chose the diverse set on 93% of the basic-level items, giving results similar to those of the oldest children in Experiment 1. For the superordinate-level items, the proportion of diverse choices was significantly lower, 69%, $t(11) = 2.93$, $p < .05$. On these superordinate-level items, the adult subjects sometimes commented that the two options were very closely balanced and that it was difficult to choose one over the other. These comments further supported the idea that the diversity effect is weaker for more remote inferences. Hence the main question behind Experiment 2 was whether children would show an adultlike pattern of restricting the use of diversity information.

Method

Experiment 2 was like Experiment 1, with the following changes. Ninety-two children, who attended Brookhurst Primary School, in Leamington Spa, participated. There were 46 students in year 1 (mean age 5;10, range 5;3 to 6;5) and 46 students in year 4 (mean age 8;10, range 8;3 to 9;5).

The stimuli for one test question from Experiment 1, relating to pets, were replaced because these stimuli belonged to a higher level taxonomic category than the other stimuli. The other stimuli belonged to basic-level categories such as balls or dolls. The photographs for the replacement stimuli were all of chocolate bars. The new basic-level target item was an Aero chocolate bar. The new nondiverse set consisted of three photographs of a man with a Milkybar. The diverse set consisted of three photographs of a man with a Twix, a Mars bar, and a Cadbury's.

In addition to the basic-level target items, each stimulus set was assigned a superordinate-level target item, as shown in Table 3. For example, for the hats, there was a basic-level target (another hat) and a superordinate-level target (a pair of shoes, also in the clothing category). Some pictures, e.g., the shoes, were used as the superordinate target for two stimulus sets. However, any subject only saw a particular picture once. The stimuli were given a random order for each subject, with the constraint that a superordinate target picture could not be used twice for the same subject.

Within each age group, half the students were given four test questions with superordinate-

TABLE 3
Target Items for Experiment 2

Basic level	Superordinate level
Rugby ball	Yo-yo
Baby doll	Yo-yo
Purple brimmed hat	Black shoes
Red top	Black shoes
White book	Newspaper
Yellow flower	Green houseplant
Blackcurrant ice cream	Crisps (potato chips)
Aero chocolate bar	Crisps

level targets followed by four test questions with basic-level targets. The other half of the students were given four basic-level target questions followed by four superordinate-level questions. We used the standard version of instructions and not the emphatic version.

Results and Discussion

As in the first experiment, we were concerned about possible carryover effects in which a child might use strategies from one question as the basis for answering a later question. Therefore we considered the first four responses from each subject to be more pure, and we initially address these data. The key result was that children made a lower proportion of diverse choices for superordinate-level target items than for basic-level target items (see Table 4). In addition, older children made more diverse choices overall than younger children. A two-way ANOVA supported these observations. There were main effects of taxonomic level, $F(1, 88) = 19.35$, $MSE = .06$, $p < .001$, and school year, $F(1, 88) = 7.41$, $MSE = .06$, $p < .01$. The interaction between these two variables did not approach statistical significance, $F < 1$.

TABLE 4
Proportion of Diverse Choices, Experiment 2

Year	Basic (*SE*)	Superordinate (*SE*)
	First four responses	
1	.71	.46
	(.05)	(.05)
4	.83	.62
	(.04)	(.06)
	Last four responses	
1	.74	.55
	(.04)	(.06)
4	.61	.64
	(.08)	(.06)

These results replicate and extend those of Experiment 1. The overall proportion of diverse choices for basic-level targets, 77%, was similar to that of the standard condition of Experiment 1, 74%. Crucially, children were less likely to make the diverse choice for superordinate-level targets, at a level of only 54%, apparently chance responding. Clearly children favored the diverse set of premises when this was most appropriate, for basic-level targets, but they did not apply this response strategy in an unconstrained way. Instead they showed the more sophisticated pattern predicted by the Osherson et al. (1990) model and demonstrated in our experiment with adult subjects, in which the preference for diverse choices is weakened for superordinate-level targets. The specific level of responding in the superordinate-level condition represents chance performance, and in this sense differs somewhat from the adult results in the pretest (69%), although adults did also report guessing in this condition. Despite this difference between children and adults, we emphasize again that the crucial prediction of weaker diversity effects for remote inferences was supported in both subject populations.

Finally, the results of Experiment 2 supported an age effect, which was also suggested by Experiment 1. However, we would hesitate to overinterpret the developmental trend because the same pattern was shown by 5-year-olds and 8-year-olds, with the older children simply showing it more strongly. Consequently, the age effect could be due to performance differences in, for example, how well children of different ages pay attention to this sort of task.

Next, turning to the last four responses, the pattern of results for year 1 students is similar to their first four responses. The slightly higher proportion of diverse responses on superordinate-level items (55% compared to 46%) could reflect a carryover effect. Having established a pattern of diverse responses on the first four, basic-level items (71%), this propensity could have carried over in some children onto the last four, superordinate-level items. For year 4 children, perhaps the surprising result is the relatively low level of diverse responses to the last four, basic-level items (61%). This result is associated with a fairly high degree of variance as well. We would attribute the apparent weakening of the diversity effect in this condition again to carryover effects. After establishing various response strategies on superordinate-level items on the first four test questions, year 4 children seemed to be slightly impaired at using diversity on the last four, basic-level items. We conclude that although children can use diversity information in inductive reasoning, even in year 4 children the effect may be susceptible to the use of other, competing strategies, a point also suggested by Gutheil and Gelman (1997). This finding also indirectly supports the view that the failure to find diversity effects in the Lopez et al. (1992) experiments might be partly due to the strategy switching necessitated by the intermixing of similarity, typicality, and diversity trials.

For completeness, we present the ANOVA based on all eight responses. Again, the key finding is a main effect of basic-level versus superordinate level, $F(1, 88) = 20.91$, $MSE = .05$, $p < .001$, with a higher proportion of diverse responses for basic-level items. This effect had a marginal interaction with year, $F(1, 88) = 3.84$, $MSE = .05$, $p < .10$, and the three-way interaction between level (basic or superordinate), year, and order (basic first or superordinate first) was significant, $F(1, 88) = 6.00$, $MSE = .05$, $p < .05$. This three-way interaction seems to reflect that older children who first saw superordinate-level items were somewhat impaired at using diversity on the last four, basic-level items. The individual variables of year and order did not have significant main effects, $F(1, 88) = 1.71$, $MSE = .10$, and, $F(1, 88) = 2.76$, $MSE = .10$, respectively. The interaction between year and interaction was also not significant $F(1, 88) = .90$, $MSE = .10$.

In summary, we would emphasize that the most important effect, the adult-like pattern of greater use of diversity for basic-level targets than for superordinate-level targets, does appear when considering all the data together, as well as on the first four, pure, items considered on their own.

EXPERIMENT 3

The first two experiments showed that children generally favored linking new target objects to the diverse set rather than the nondiverse set. The third experiment was aimed not so much at the effect of the diverse set but rather the nature of the nondiverse set. For example, when children saw three pictures of Danielle playing with a Barbie doll, the situation was somewhat ambiguous. Did Danielle have three Barbie dolls? Or was she playing with the same Barbie doll on three different occasions? That is, were the three dolls genuinely perceived to be three different tokens of the same type, or were they seen as three pictures of the same token? In line with a diversity task, the spoken instructions referred to the second and third dolls as ''another Barbie doll,'' thus encouraging an interpretation in terms of three distinct tokens. But this interpretation cannot simply be assumed. Although Danielle and the doll were posed differently in each photograph, the photographs were in fact of the same Barbie doll three times (in contrast, it was necessarily clear from both the instructions and the photographs that Jane had three different dolls in the diverse condition).

This argument raises the possibility that the results of the first two experiments are more of a sample size effect rather than a diversity effect, at least for children who scrutinized the pictures very carefully. When deciding who the new doll belongs to, children might have reasoned that Jane has more dolls than Danielle rather than that Jane has a more diverse set of dolls than Danielle. Even a sample size effect in children would be an important finding. Lopez et al. (1992) and Gutheil and Gelman (1997) reported mixed results on whether 9-year-olds could use sample size in inductive reasoning.

For stimulus sets with specific conclusion categories, Lopez et al. reported that 9-year-olds did not show sample size effects. In general, Gutheil and Gelman did not find sample size effects in 9-year-olds either, except in limited circumstances. Lopez et al. (1992) also reported that 5-year-olds failed to show sample size effects. So even if Experiments 1 and 2 were merely showing sample size effects in children age 5 to 9, these would still be important and novel findings. We note, however, that the pattern of results in Experiment 2, showing differences in reasoning about basic-level matches compared to superordinate-level matches, already seems to go somewhat beyond sample size effects; there seems to be a more subtle pattern of results.

For Experiment 3, we removed any possible ambiguity on the sample-size versus diversity issue. We used the same diverse sets of pictures from Experiment 2, but we reshot the photographs of the nondiverse sets. For example, the nondiverse set for dolls was three pictures of Danielle playing with different Barbie dolls (differing in hair and clothes). Similarly, the nondiverse set of balls clearly showed Tim playing with three different soccer balls (different colors and markings). We expected that children would continue to show diversity effects with these stimuli.

Method

Experiment 3 was like Experiment 2, with the following changes. Fifty-seven children participated, from Cubbington Primary School in Leamington Spa. There were 25 students in year 1 (mean age 6;4, range 5;11 to 6;10) and 32 students in year 4 (mean age 9;4, range 8;11 to 9;9).

For stimuli, the diverse sets from Experiment 2 were used again, and likewise the basic-level target items were used. The nondiverse sets of photographs from Experiment 2 were replaced, with the new photographs making it clear that there were three instances of each type, such as three different soccer balls. Only the standard version of instructions and not the emphatic version was used.

Results and Discussion

Overall, as shown in Table 5, children robustly favored the diverse choice over the nondiverse choice. The overall proportion of diverse choices, .81, was actually higher than the comparable results in the first two experiments. This value was also significantly greater than a chance level of 50%, $t(56) = 11.56$, $p < .001$. Inspection of Table 5 suggests that this tendency was even

TABLE 5
Proportion of Diverse Choices, Experiment 3

Year	Mean (*SE*)
1	.73 (.04)
4	.87 (.03)

stronger for older children. An ANOVA indicated a significant effect of year in school, $F(1, 55) = 7.66$, $p < .01$, $MSE = .04$.

These results showed unambiguous diversity effects for children in year 1 and year 4. In this experiment, it is not possible to attribute the results to a sample size effect. Although it may be possible to attribute the results of Experiments 1 and 2 to a mixture of diversity effects and sample size effects, given the strong results of Experiment 3 it is clear that the diversity effect is no weaker than the sample size effect.

Like Experiment 1, with its strong diversity effects, Experiment 3 was well-suited to examination of individual items because a large amount of data was collected for each item, without breaking down the experiment into subconditions. Again, we found that diversity effect was consistent across items, with the proportion of diverse choices ranging from 70 to 88% across stimulus sets. Therefore, the results favoring diversity did not seem to depend on the idiosyncrasies of a small number of stimuli.

Also, there was a sizeable enough pool of data to analyze the explanations given by children after the eighth trial. Of the 57 children, 35 gave diversity-based explanations, coded using the same criteria as in Experiment 1. For 8 children, no explanations were recorded. The remaining 14 children gave other explanations which were unclassifiable or similarity-based. Half of these children chose the nondiverse set and half chose the diverse set on the last trial. Although the pattern of explanations is slightly different than Experiment 1, the key findings were repeated. The predominant form of explanation was diversity-based. There was some evidence for similarity-based reasoning as well, but the overall result favoring diverse choices seemed to be mainly attributable to diversity-based reasoning rather than to responding to similarities between the target item and the given items.

Our focus in Experiment 3 on the role of the nondiverse set raises a general question about research on diversity effects. When people draw stronger inferences from a diverse set of evidence compared to a nondiverse set of evidence, are they responding to the diversity of one set, to the nondiversity of the other set, or to the difference in diversity between the two sets? Again, we see this as a general question that can be raised about many studies of diversity rather than an issue that our experiments were specifically designed to address. Nonetheless, some of our results seem to bear on this issue indirectly. In particular, it does not seem that children were simply noticing that one stimulus set is not diverse then consequently rejecting it. For example, it does not seem that children were merely noting that Danielle always played with the same kind of doll and therefore responding that Jane would play with new items. In Experiment 2, we kept the nondiverse set constant, and varied the target item (such as another doll or a ball). If children were simply noting that one stimulus set is nondiverse, that would not explain the weaker diversity effects for remote target items. Likewise, in Experiment 3, we increased the diversity of the nondiverse sets somewhat by, for example, using

three soccer balls with different markings compared to the same markings in Experiments 1 and 2. We still obtained robust diversity effects in Experiment 3, even after making the nondiverse sets less uniform. Informally, we noted that some of the children's explanations referred to the diversity of one set and other explanations referred to the nondiversity of the other set. In sum, although our experiments were not specifically aimed at this issue, it seems unlikely that children were merely noticing that one of the stimulus sets was nondiverse.

EXPERIMENT 4

Together, the first three experiments documented that children in the age range of 5 to 9 years can reason on the basis of diversity in an inductive reasoning task. In a sense, the results stand on their own as empirical findings, without any need to compare to other studies. Whether our first three experiments perfectly map onto past studies is in some ways a secondary issue. Just showing that children do show robust diversity effects has implications for any model or account of inductive reasoning that addresses diversity. In coming up with a model that can address reasoning by adults and children, establishing that both do show diversity effects suggests that the same mechanisms could be involved.

However, it is natural to go further and consider why the children in our experiments did indeed show such robust diversity effects, or alternately why children in past studies (Carey, 1985, Rozelle et al., 1999; and especially Lopez et al., 1992; and Gutheil & Gelman, 1997) did not show diversity effects. Given our results that children are capable of diversity-based reasoning, we are led to consider issues of performance and knowledge in explaining the different results in past studies, rather than considering differences between children's reasoning processes and adults' reasoning processes. Of course, a comprehensive investigation of all the variables that might promote or prevent the use of diversity information is beyond the scope of this article. Such an investigation would likely require a long series of experiments. However, we begin this process here by focusing on some of the similarities and differences between our work and past studies. Experiment 4 was conducted with the aim of looking at some of the potentially most important variables.

As already stated, the key similarity between the past studies, particularly Lopez et al. (1992) and Gutheil and Gelman (1997), and our own experiments is the central design. In each experiment, there is a diverse set of items and a nondiverse set of items. The child learns one fact for each set, and then the child is asked which fact would apply to an additional target item. In addition, there were several differences between our experiments and the past studies. Some of the main differences start to arise when considering what the items and facts are for each experiment. For the past studies, the

items were animals, and the facts were hidden internal properties, typically of a biological or anatomical nature. For our first three experiments, the items were everyday objects such as dolls and balls and hats. The facts were which person went with each object. It is immediately clear that there are content differences—we did not refer to biological facts although some of our stimulus items were living things such as cats and flowers. Instead we used everyday objects and facts. (See Heit & Rubinstein, 1994, for additional examples of content effects.) Two other differences are that in past studies the facts referred to properties, whereas in our first three experiments the facts referred to relations; the properties were hidden in past studies but the relations were visible in our experiments.

Another difference has to do with wording. The past experiments used object-first wording, such as ''Cats have leukocytes inside.'' Our first three experiments used object-last wording, such as ''This is Tim. He's playing with a football.'' Though we did not expect that word order would have a dramatic effect on the results, we thought it would be valuable to try to rule out word order as having a negative influence on performance by using an object-first word order, as in previous research.

Thus, Experiment 4 had two main aims. One was to make the wording of our instructions more like the instructions of past studies by naming the object first rather than the person first. For example, the experimenter said ''This chocolate bar belongs to Theresa'' rather than ''This is Theresa. She's eating a chocolate bar.'' The other aim was to compare two kinds of facts. In the person condition, the stimuli referred to an object in some visible relation with a person, such as the chocolate bar and Theresa. In the insides condition, the stimuli referred to an object with some hidden property, such as ''This chocolate bar has nuts inside.'' The insides condition was meant to be more like past studies, which also referred to hidden properties, such as a cat with leukocytes inside. However, Experiment 4, like our first three experiments, stayed within the content domain of everyday objects and did not involve biological knowledge as in past studies.

Method

Experiment 4 was similar to the previous experiments, with the following changes. Ninety-two children participated, from Cannon Park Primary School in Coventry, a city in the midlands of England. There were 46 students in year 1 (mean age 5;8, range 5;2 to 6;2) and 46 students in year 4 (mean age 8;8, range 8;2 to 9;1). Within each year, the students were randomly assigned into two equal-sized groups, corresponding to the person condition and the insides condition.

New stimulus sets were created for this experiment, as shown in Table 6. As in the previous experiments, there was always a nondiverse set of objects and a diverse set of objects, with an additional target item. In addition, two hidden properties were generated for each test question. These two properties could potentially apply to the target item as well as any of the objects in the nondiverse and diverse sets for this question. For example, with the bag stimuli, any bag could contain jelly babies or dolly mixtures (two kinds of candy). Likewise, any of the dolls could be wearing shoes or not wearing shoes. (The dolls' feet were not photographed.)

TABLE 6
Stimuli for Experiment 4

Target item	Nondiverse set	Diverse set	Hidden properties
Purple (wrapped) chocolate bar	Green chocolate bar	Silver chocolate bar, orange chocolate bar, gold chocolate bar	Nuts inside, raisins inside
Green bag	White bag	Pink bag, yellow bag, blue bag	Jelly babies inside, Dolly mixtures inside
Yellow book	Red book	Green book, blue book, purple book	Pictures of Teletubbies inside, pictures of Spice Girls inside
Stuffed (toy) fish	Stuffed bird	Stuffed pig, stuffed rabbit, stuffed dog	Makes a rattling sound, makes a squeaking sound
Red drink	Yellow drink	Green drink, blue drink, purple drink	Tastes sweet, tastes sour
Green hat	Blue hat	Black hat, brown hat, red hat	Furry inside, silky inside
Yellow flower	Pink flower	Purple flower, blue flower, green flower	Goes in a big flower pot, goes in a small flower pot
Clown doll	Plastic doll	Baby doll, boy doll, Barbie doll	Wearing shoes, not wearing shoes

Note. These descriptions of objects are summaries. Different objects tended to differ in size, shape, and coloring.

In the person condition, each set of photographed objects was displayed next to a person. For example, three photographs of green wrapped chocolate bars would be shown next to a photograph of Theresa. The photographs of silver, orange, and gold chocolate bars would be set next to a photograph of Linda. This method was slightly different than the previous experiments, in which the person interacted with the object in each photograph, e.g., there were three photographs of Tim with a soccer ball. However, having separate photographs of the objects allowed us to use the same photographs for the insides condition. In this condition, the same sets of photographs were shown, without the pictures of people.

The spoken presentation was the same as our other experiments, except that the name of the object was always mentioned before its associated person or property. The wordings for the person and insides conditions were similar except that in the person condition, the child was told that each object belonged to some person, whereas for the insides condition, the child was told that each object had a particular hidden property. The first four items were presented with standard instructions and the last four items were presented with emphatic instructions.

Results and Discussion

Overall, as shown in Table 7, the results in the person condition for standard and emphatic instructions were roughly similar to the results of Experiment 1 in Table 2. Children in both age groups tended to make the diverse

TABLE 7
Proportion of Diverse Choices, Experiment 4

	Person		Insides	
Year	Standard (*SE*)	Emphatic (*SE*)	Standard (*SE*)	Emphatic (*SE*)
1	.63 (.06)	.79 (.05)	.54 (.05)	.59 (.05)
4	.88 (.04)	.91 (.03)	.66 (.06)	.77 (.06)

choice for both sets of instructions. Therefore, based just on the person condition, it appears that word order does not make much of a difference. Children show comparable diversity effects for object-first word order (Experiment 4) as well as object-last word order (Experiment 1).

For the insides condition of Experiment 4, the proportions of diverse choices are generally lower than the person condition. A three-way ANOVA was conducted, with condition (person or insides), year (1 or 4), and instructions (standard or emphatic) as variables. There was a main effect of condition, $F(1, 88) = 15.08$, $MSE = .08$, $p < .001$, supporting the observation that there were more diverse choices in the person condition. There was also a main effect of year, $F(1, 88) = 16.10$, $MSE = .08$, $p < .001$, with older children making more diverse choices, and a main effect of instructions, $F(1, 88) = 9.92$, $MSE = .04$, $p < .01$, with emphatic instructions leading to more diverse choices. None of the interactions between these variables reached the level of statistical significance, although the three-way interaction was marginal, $F(1, 88) = 3.14$, $MSE = .04$, $p < .10$, possibly reflecting a ceiling effect in the person condition for year-4 children.

Consistent with the findings from the ANOVA, Table 7 shows that the lowest proportion of diverse responses was by made year-1 children in the insides condition, with standard and emphatic instructions. As a rough guide to where diversity effects were evidenced, neither of these cell means (.54 and .59) were significantly different from chance using a *t* test and a $p < .05$ criterion; however, the remaining cells mean in Table 7 did differ significantly from 50%. Likewise, when pooled together the results for year-1 children in the insides condition did not significantly differ from chance, but in comparison, the proportions of diverse choices for year-1 children were significantly greater than chance for the person stimuli in Experiments 1, 2, and 3 (except for the superordinate condition of Experiment 2).

We conclude from Experiment 4 that using external relations to a person does promote diversity-based reasoning relative to using hidden internal properties. This comparison was made using the same set of objects in both cases, so the difference must be due to the facts about the objects rather than the objects themselves. Why might using visible relations enable an adultlike

pattern of reasoning compared to hidden properties? We see two components of the advantage of visible relations. First, it is plausible that visible information especially provides support to cognitive processes that are still very sensitive to disruption. A different example for the boost external representations can give to children's performance was given by Mitchell and Lacohee (1991). They showed that 3-year-olds' performance on a theory of mind task was enhanced by having them draw a picture of their initial belief states. However, our Experiment 4 actually used somewhat less visible information than our previous experiments because the person and the object were not shown interacting. Even with this reduction of visibility, we still obtained robust diversity effects.

Second, there are many results suggesting that children are relatively at ease considering thematic relations between objects, i.e., which object goes with what (e.g., Markman, 1981; Markman, Horton, & McLanahan, 1980). A relation such as a person playing with an object or wearing an object can be considered a thematic relation. In addition, in the person condition of Experiment 4 as well in as our other experiments, the objects themselves seem to take on another thematic relation, that of a collection. We are seeing Theresa's collection of chocolates and Linda's collection of chocolates. In contrast, just seeing three chocolate bars with nuts inside does not make a collection of chocolate bars; it is just three chocolate bars with a common property. Hence, it is plausible that both the visible nature of the relations, as well as the use of relations rather than properties, contributed to the greater use of diversity-based reasoning in the person condition compared to the insides condition.

In addition, is it plausible that the particular visible, external relation we used, possession of objects by people, could have facilitated a systematic pattern of inference due to children's familiarity with reasoning about people and their possessions. In comparison, some of the hidden, internal properties we used may not have seemed quite so predictable according to children's prior experience. Although the wrapper of a chocolate bar is sometimes a good predictor of what is inside, the color of a book's cover is often not a good predictor of what is inside. However, even using hidden properties, we obtained statistically significant evidence for diversity effects in 8-year-olds. This result in itself is novel because the past studies that also used hidden, internal properties and specific conclusions did not find diversity effects in 9-year-olds. In the comparable condition for 5-year-olds, the proportion of diverse responses was not significantly above chance. It would be important in future experiments to investigate whether other factors could facilitate performance by 5-year-olds using hidden, internal properties or if there is some fixed limitation to their abilities.

Finally, the difference between the person condition and the insides condition for the 5-year-olds suggests that they were not simply responding to accidental similarities in the stimuli between the target item and the given

items. If the robust diversity effects in the person condition just reflected similarity, then surely there would be comparable effects in the insides condition—Lopez et al. (1992) have already documented robust similarity effects in 5-year-olds using hidden properties.

GENERAL DISCUSSION

In contrast to past studies, our results show clearly that children from age 5 to age 9 can perform diversity-based reasoning, using familiar categories. In addition, this diversity-based reasoning is sophisticated enough to be sensitive to the scope of the inference, with children showing attenuated diversity effects for remote inferences about more distantly related target categories. The results do not seem to be due to a sample size effect rather than a diversity effect. Although the word order in the first three experiments was different from past studies, the fourth experiment also showed diversity effects with the alternate word order. We did find that using hidden, internal properties reduced the proportion of diversity-based choices compared to visible external relations to people.

As we noted at the start of this article, similarity between premise and conclusion categories has been shown to be a dominant influence on inductive inference (e.g., Rips, 1975). There is substantial evidence in our experiments that children were not simply responding to similarities between premise statements and conclusion statements when they showed diversity effects, but they were actually assessing and responding to the diversity of the premise statements. For example, children tended to say that Robby, who played with a basketball, cricket ball, and tennis ball, would also play with a rugby ball, rather than Tim, who played with three soccer balls. Although this result might be due to some accidental similarity, such as a chance resemblance between the basketball and the rugby ball, we attributed the results to the diversity of Robby's possessions relative to Tim's possessions.

The first line of evidence that we truly obtained diversity effects rather than inadvertent similarity effects is that the results went in the same direction for different items. In Experiments 1 and 3 we collected the most data per item, so these experiments are appropriate for splitting up the results by item. Across the 16 item sets in these two experiments, the lowest proportion of diverse choices was 69%. It did not seem that accidental similarities in a few item sets distorted the overall pattern of results—the overall pattern of diversity effects was mirrored in each of the individual item sets. Furthermore, in Experiments 1 and 4, we found that emphasizing the diversity and nondiversity of item sets lead to an increase in the diversity effect. If the results were merely due to premise-conclusion similarities rather than diversity, it is hard to see why emphasizing diversity would strengthen the results.

The third source of evidence that our results were not merely due to similarity is the chance results in the internal property condition of Experiment

4, for 5-year-olds. In contrast, Lopez et al. (1992) showed robust similarity effects in 5-year-olds, using hidden internal properties. If our experimental stimuli so readily promoted similarity effects, then they would be expected to do so for the young children in Experiment 4 as well. Finally, the majority of children's explanations, recorded on the final test trials in Experiments 1 and 3, explicitly mentioned diversity or nondiversity of the premise items. To the extent that children mentioned similarity between premise and conclusion items, these explanations were associated with making the nondiverse choice rather than the diverse choice. (Indeed, Gutheil and Gelman, 1997, found that preventing children's use of premise–conclusion similarity actually increased the use of diversity somewhat.) In sum, our results would be difficult to explain in terms of accidental similarities between premise and conclusion categories, and they are readily explained in terms of diversity of premise categories.

Therefore we would conclude that in terms of diversity effects, children can assess evidence and reason about categories in a manner roughly similar to adults, when conditions permit. Their reasoning with multiple categories or multiple sources of evidence does not seem to be deficient compared to adults, provided that the materials being reasoned about and the other task requirements are familiar and manageable enough.

Historically, one of the central questions in the study of inductive reasoning has been what makes some properties or facts easier to project or infer than others. For example, Gelman (1988) compared stable, internal properties to more transient or idiosyncratic properties, with children as young as age 4. For properties such as ''has pectin inside,'' children's inferences showed robust similarity effects. But for properties such as ''has a little scratch on it,'' children showed chance patterns of reasoning. Goodman (1955) referred to the issue of what makes a predicate suitable for inference as the riddle of induction, with his famous example that people would readily draw inferences about the property ''green'' but not about the property ''grue,'' where grue is defined as green before the year 2000 and blue after the year 2000. Goodman argued that some predicates are better entrenched than others. In brief, past successful use of a predicate promotes its future use in induction. (See Shipley, 1993, and Ellison, 2001, for further discussion.)

Note that even an unfamiliar predicate can be very entrenched because it can inherit entrenchment from its class. For example, the predicate ''belongs to Jane'' was in a sense unfamiliar to the children in our experiments because they had never seen Jane before. However, the general class of predicates that some object X ''belongs to person Y'' would be extremely familiar to children. Numerous everyday transactions and interactions in children's lives involve attending to which objects belong to which people and acting appropriately based on this information. Hence, we would expect the external relations taught in our experiments, that particular objects belonged to particular people, to be extremely entrenched for children, and most likely to show a

robust and systematic pattern of inductive inferences. In comparison, predicates relating to internal organs and body chemistry of animals would have less entrenchment for children and would not give as good support for systematic inferences.

To give a broader view of the developmental results, we now discuss their relation to cross-cultural research on diversity effects as well as implications for both descriptive and normative models of induction.

Relations to Cross-Cultural Research

Looking at the broad range of developmental studies on diversity effects, it is clear that there are some situations where the phenomenon occurs consistently as well as some places where the phenomenon consistently does not occur. Likewise, recent work on diversity effects in adult populations other than American college students has provided a variety of fascinating results. Whereas Choi, Nisbett, and Smith (1997) reported that Korean college students showed diversity effects, for both animal categories and categories of people, and likewise Viale and Osherson (2000) reported that Italian college students showed diversity effects, there have been some well-documented exceptions with adults. In their study of Itzaj-Mayan adults from the rainforests of Guatemala, Lopez, Atran, Coley, Medin, and Smith (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. Based on 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 get some disease, one subject claimed it would be easy for the shorter kinds of palm trees, located below, to get the disease as well.

In a follow-up study, Lopez et al. (1997) found that the Itzaj do show diversity 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.) Subjects tended to prefer the latter, more diverse choice. Therefore, the Itjaz seemed to follow the principle of diversity unless some other source of knowledge suggested a different pattern of inference.

Giving further support to this idea that other strategies and knowledge can overrule diversity, Proffitt, Coley, and Medin (2000) have recently reported that American adults who are tree experts (such as landscapers and park maintenance workers) did not show strong diversity effects when reasoning about trees and their diseases. The tree experts seemed to be relying on the knowledge that tree diseases tend to spread readily within tree families such as elm and maple. Their inferences seemed to follow an alternate strategy

that did not assess diversity against the broad category of ''all trees'' but rather considered the size of various tree families. (There are related results in Carey, 1985, in which children seemed to rely on subcategories such as mammal and insect without considering the broad category of ''all animals.'') Coley, Medin, Proffitt, and Lynch (1999) discussed both the Itzaj and tree expert results further, putting together a strong case for how prior knowledge can overcome diversity effects.

There are especially intriguing parallels between the Lopez et al. (1997) results and the developmental results. Just as Lopez et al. found diversity effects in Itzaj adults with one version of the task but not in a version with based on different materials, we found diversity effects in 5-year-olds with one set of materials, whereas past researchers did not find diversity effects with other materials. Likewise, we found differences within Experiment 4 in a direct comparison of two kinds of materials. Just as the lack of diversity effects in the Itzaj seemed to be due to their great expertise with some stimulus materials, the lack of diversity effects in children may be attributable, at least in part, to discrepancies between the experimenters' and children's expertise.

Whereas Western schoolchildren, particularly older children, clearly possess biological knowledge both in the form of knowledge of a wide variety of animals and some rudimentary knowledge of their properties, children's base of relevant knowledge clearly differs sufficiently from that of adults to leave ample room for discrepancies to emerge. It seems likely that particularly younger children, whose biological world is heavily influenced by the fictitious materials of storybooks, films, and toys, would engage in knowledge-based inferences which to adults would seem alien. Furthermore, predicates relating to internal organs and the body chemistry of animals would seem to have less entrenchment for children and would not give as good support for systematic inferences. Finally, children's use of diversity could be fairly fragile and thus easily disrupted when other information or strategies are available. Therefore, the use of diversity could be masked or counteracted when different types of questions are intermixed, for example, in Lopez et al. (1992) and in our own Experiment 2. In general, what may be developing in children is not the ability to use diversity information but rather the ability to coordinate different strategies and pools of knowledge and to use them consistently in inductive inference.

Implications for Models of Induction

Based on their results showing a lack of diversity effects in children, Lopez et al. (1992) concluded that models of inductive reasoning by adults would need to be modified to explain reasoning by children. In the terms of the influential Osherson et al. (1990) model, it was suggested that children may have difficulty assessing diversity because their mechanism for generating inclusive categories is not developmentally ready. Gutheil and Gelman

(1997) disputed some of the details of this conclusion, but the general issue remains whether the same formal account of induction can be applied to children and adults. (See Heit, 2000, for a review of other models of induction and how they address diversity effects.)

In light of our own results showing diversity effects in children, as well as the compelling evidence for use of other knowledge and strategies in cross-cultural work with adults, we see the crucial issue to be captured by modeling not as developmental change in mechanisms of induction but rather the influence of knowledge on induction. Putting together our results with past studies, it is clear that there are dramatic content-related influences on inductive reasoning. Whether diversity effects appear depends on the relation between what children and adults are reasoning about and their general knowledge. Reasoning about everyday objects with people leads to different outcomes than reasoning about animals and their internal biology or other hidden properties.

As reviewed by Heit (2000), extant models of inductive reasoning need to give a better account of how prior knowledge influences induction to explain inductive reasoning by adults as well as children. A model that always predicts stronger inferences for more diverse sets, regardless of the content or domain, is not going to be able to account for inductive reasoning by children or any subject population for that matter. Likewise, to the extent that our results are attributed to differences between external, visible relations and internal, invisible properties, a model of reasoning would need to address these differences as well. One particular question that would need to be addressed by future modeling work is whether the appearance versus nonappearance of diversity effects in children and adults can be attributed to a common explanation. Perhaps for both age groups, the nonappearance of diversity effects in some situations is best explained in terms of other sources of knowledge overriding a general strategy to use diversity information. Alternately, the lack of diversity effects in children might be better addressed with a different explanation such as unsystematic reasoning due to a lack of knowledge about and relatively low entrenchment of biological predicates.

Normative Analyses

In addition to work on descriptive models of inductive reasoning there has been a tradition of looking at induction from a normative perspective, including the standard claim that diverse evidence is stronger evidence. In general, we see the potential for a two-way interchange between normative accounts of induction and results of psychological experiments on induction. Normative accounts can explain, for example, why it would be adaptive to draw stronger inferences from diverse observations than from nondiverse observations. When there are surprising empirical exceptions to normative predictions, such as 9-year-olds not showing diversity effects, such results might suggest that further experimentation is needed, for example, to un-

cover children's true potential. Empirical exceptions to normative predictions can also be an impetus to developing better normative accounts. For example, exceptions to diversity effects with experts in the domain of living things (Lopez et al., 1997; Proffitt et al., 2000) suggest that normative accounts might be broadened to encompass other sources of information such as expert knowledge. We now briefly review the history of normative arguments for why diverse evidence is strong evidence before pointing to some ways that normative accounts would need to better address psychological phenomena.

The value of diverse evidence for testing a hypothesis has been stressed repeatedly in the philosophical literature on scientific reasoning. (See also Lopez, 1993, and Spellman, Lopez, & Smith, 1999, for further discussion.) The earliest argument we have found for the value of diverse observations is in Bacon's *Novum Organum* (1620/1898), which cautioned against hasty generalizations drawn from narrow samples. More recently, Nagel (1939) argued, on probabilistic grounds, that establishing a scientific theory should use diverse observations rather than a lot of similar observations to obtain more reliable estimates. He gave the example of inspecting the quality of coffee beans delivered on a large ship. It would be better to inspect small samples of beans from various parts of the cargo hold than to inspect a large number of beans from just one part of the ship. Carnap (1950) addressed the problem of why simply collecting larger numbers of observations does not always lead to greater confidence. Eventually, collecting increasing numbers of highly similar observations should not give much extra evidence for a scientific theory. He linked the collection of diverse evidence to the desirable quality of scientific theories that they should make novel predictions rather than merely redescribe old data. A scientific theory should be strongly supported if it makes diverse predictions that are subsequently verified. Similarly, Hempel (1966) related the collection of diverse evidence to a falsifying research strategy, which he argued provides better support for scientific theories than a confirmatory research strategy.

These intuitions have led to several attempts, typically in terms of Bayesian statistics, to prove or formalize the evidential advantage obtained from diverse evidence. As reviewed by Wayne (1995), there have been two main lines of approach. The first approach compares correlated sources of evidence to independent sources of evidence. Due to the similarity between lions and tigers, knowing that lions have ulnar arteries makes it seem likely that tigers have ulnar arteries as well. Once a person has observed that lions have ulnar arteries, observing that tigers have ulnar arteries does not seem to add much independent or surprising information. In contrast, observing that giraffes have ulnar arteries would seem less predictable, due to the lower similarity between lions and giraffes. Hence the combination of lions and giraffes provides stronger evidence to promote further inferences. For formal

treatments of this approach, linking similarity to probability theory, see, e.g., Earman (1992), Howson and Urbach (1993), and Viale and Osherson (2000).

The second approach is the eliminative approach. The idea behind the eliminative approach is that diverse data sets will be particularly useful for eliminating plausible but incorrect hypotheses, allowing stronger inferences to be drawn based on the remaining, contending hypotheses. In contrast, nondiverse data sets will likely be consistent with too many hypotheses to allow any strong inferences. Horwich (1982) gave a proof based on the abstract example of trying to extrapolate from three points that fall on a straight line. If the three points are far apart, Horwich argued that it is reasonable to extrapolate along a straight line, rather than fit some other, nonlinear functions to the points. On the other hand, if the three points are bunched closely together, it is easy to entertain many alternative hypotheses for how to extrapolate from the three points, and no particular conclusion will be strongly indicated.

How can these normative accounts be better linked to the psychological results on diversity? One preliminary point is that to apply any of these formal analyses to inferences about specific conclusion categories, such as dogs rather than all mammals, additional assumptions would be needed, such as an additional step of deductive inference (e.g., if all mammals have some property then dogs must have the property). But more crucially, how can these accounts be expanded to address not only the general pattern of diversity effects but also the systematic exceptions to diversity? Induction is by definition a case of uncertain inference, which should take into account further background knowledge when available. Any normative, probabilistic account of induction should be able to incorporate background knowledge. This contrasts sharply with the case of deduction where the inclusion of background knowledge typically constitutes a fallacious intrusion (for example, belief bias effects in conditional inference; Evans, Newstead, & Byrne, 1993).

We see the development of normative accounts of induction to incorporate both a structural bias in favor of diverse evidence as well as information derived from other background beliefs to be an important area of future work. Some recent model developments have started to move in this direction. Working from the general framework of probability theory and specifically the idea that more surprising observations should lead to greater belief revision, Rozelle et al. (1999) have suggested that other sources of prior knowledge could make evidence seem surprising. Hence even nondiverse evidence can lead to strong inferences, if this evidence is surprising in light of other knowledge. Taking an alternate approach, Heit (1998, 2000) has proposed a Bayesian model of inductive reasoning that is much in the spirit of Horwich's (1982) account of induction in terms of eliminating hypotheses. This Bayesian model was applied to a number of psychological phenomena such as

those in Rips (1975) and Osherson et al. (1990). Heit (1998) described several ways in which prior beliefs can be represented in terms of the initial hypothesis space, leading to various knowledge effects on inductive reasoning and in some cases producing inferences that would on the surface appear to be nonnormative.

In sum, taking account of the diversity of observations, and favoring diverse sources of evidence, seems to promote more accurate estimation, broader generalizations, and stronger theories of the world, especially when put together with other sources of knowledge. Having such a powerful tool available early on in development thus seems of great adaptive value.

REFERENCES

Bacon, F. (1620–1898). *Novum organum.* London: George Bell and Sons.

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

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

Choi, I., Nisbett, R. E., & Smith, E. E. (1998). Culture, category salience, and inductive reasoning. *Cognition,* **65,** 15–32.

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

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory.* Cambridge, MA: MIT Press.

Ellison, T. M. (2001). Induction and inherent similarity. In U. Hahn & M. C. A. Ramscar (Eds.), *Similarity and categorization.* Oxford, UK: Oxford Univ. Press.

Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction.* London: Erlbaum.

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

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

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

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.

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, UK: Oxford Univ. Press.

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

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). Hillsdale, NJ: Erlbaum.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and 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 and Memory,* **2,** 322–330.

Horwich, P. (1982). *Probability and evidence.* Cambridge, UK: Cambridge Univ. Press.

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach.* Chicago: Open Court.

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

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

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

Mareschal, D., French, R. M., & Quinn, P. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology,* **36,** 635–645.

Markman, E. M. (1981). Two different principles of conceptual organization. In M. E. Lamb & A. L. Brown (Eds.), *Advances in developmental psychology.* Hillsdale, NJ: Erlbaum.

Markman, E. M., Horton, M. S., & McLanahan, A. G. (1980). Classes and collections: Principles of organization in the learning of hierarchical relations. *Cognition,* **8,** 227–241.

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

Mitchell, P., & Lacohee, H. (1991). Children's early understanding of false belief. *Cognition,* **39,** 107–27.

Nagel, E. (1939). *Principles of the theory of probability.* Univ. of Chicago Press.

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

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

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

Quinn, P. C., Eimas, P. D., & Rosencrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception,* **22,** 463–475.

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

Rozelle, J., Sides, A., & Osherson, D. (1999). Evidential diversity and premise probability in young children's inductive judgment. Unpublished manuscript.

Shipley, E. F. (1993). Categories, hierarchies, and induction. In D. L. Medin (Ed.), *Psychology of learning and motivation* (Vol. 30, pp. 265–301). Orlando, FL: Academic Press.

Spellman, B. A., Lopez, A., & Smith, E. E. (1999). Hypothesis testing: Strategy selection for generalising versus limiting hypotheses. *Thinking and Reasoning,* **5,** 67–91.

Viale, R., & Osherson, D. (2000). The diversity principle and the little scientist hypothesis. *Foundations of Science,* **5,** 239–253.

Wayne, A. (1995). Bayesianism and diverse evidence. *Philosophy of Science,* **62,** 111–121.

Younger, B. A. (1985). The segregation of items into categories by ten-month-old infants. *Child Development,* **56,** 1574–1583.

(Accepted January 22, 2001; published online May 29, 2001)