Diversity-Based Reasoning in Children Age 5 to 8

Evan Heit (E.Heit@warwick.ac.uk)
Department of Psychology, University of Warwick
Coventry CV4 7AL, United Kingdom

Ulrike Hahn (HahnU@cardiff.ac.uk)
School of Psychology, Cardiff University
Cardiff CF1 3YG, Wales, United Kingdom

Abstract

One of the hallmarks of inductive reasoning by adults is the diversity effect, namely that subjects draw stronger inferences from a diverse set of premise statements than from a homogenous set of premises (Osherson et al., 1990). However, past developmental work (Lopez et al., 1992; Gutheil & Gelman, 1997) has not found diversity effects with children age 9 and younger. In our own experiments, we found robust and appropriate use of diversity information in children as young as 5 years. For stimuli we used pictures of people and their possessions, rather than the stimuli concerning animals and their biological properties in past studies. We discuss implications of these results for models of inductive reasoning.

Introduction

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 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 towards similarity-based reasoning, namely people were highly sensitive to the similarity between the given, or premise, category, and the target, or 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.

A limitation of this early work is that it only looked at inferences from one category to another. 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 processes underlying inductive ability. The most extensive and influential work on induction from multiple categories was conducted by Osherson, Smith, Wilkie, Lopez, and Shafir (1990). They reported several phenomena involving reasoning with multiple premise categories, but we will 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, below the line.

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

Rabbits have an ulnar artery.

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

Rabbits have an ulnar artery.

People 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 non-diverse 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 category-based component, in which people generate a superordinate category and assess how well the premise categories cover this superordinate. In the present example, lions and giraffes would cover the superordinate, 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 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 (1993) devised a stricter test of diversity-based reasoning, in which people chose premise categories rather than simply evaluating arguments given a set of premises. In other words, will peo-

ple's choices of premises reveal that they value diverse evidence? Subjects (American college students, as in Osherson et al.) were given a fact about one mammal category, and they 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 (1993) that for inductive arguments about animals, subjects do make robust use of diversity in not only evaluating evidence but also in seeking evidence. The results are less clear for other subject populations, however. In particular, developmental work has generally failed to find diversity effects in children. (See also Lopez, Atran, Coley, Medin, & Smith, 1997, and Choi, Nisbett, & Smith, 1997, for cross-cultural results.)

The first study of diversity-based reasoning was a developmental one by Carey (1985), comparing 6 year olds and adults. Carey looked at patterns of inductive projection given the premises that two diverse animals, dogs and bees, have some biological property. The purpose of this study was to see whether subjects reason that "if two such disparate animals as dogs and bees" have this property then "all complex animals must" (p. 141). Indeed, adults made broad inferences to all animals, extending the property not only to things that were close to the premises (other mammals and insects) but also to other members of the animal category (such as birds and worms). In contrast, the children seemed to treat each premise separately; they drew inferences to close matches such as other mammals and insects, but they did not use the diversity information to draw a more general conclusion about animals. Therefore in this first attempt there was evidence for effects of diversity in adults but not children. In a follow-up study, Carey looked at diversity effects based on the concept of living thing rather than animal. The results were actually less clear for this study, but again, there was not definitive evidence for mature diversity-based reasoning in children.

Continuing along this line of looking for diversity effects in children, Lopez, Gelman, Gutheil, and Smith (1992) found limited evidence for 9 year olds and no evidence for 5 year olds. For the 5 year olds, choices in a picture-based task did not show any sensitivity to diversity of premise categories, even when the diversity was emphasized by the experimenter. However, 9 year olds did show sensitivity to diversity of premises, but only for arguments with a general conclusion category such as animal. They did not show diversity effects at all for a specific conclusion category such as rabbit. Therefore it seemed that diversity-based reasoning was somewhat shaky in 9 year olds and not at all present in 5 year olds. Lopez et al. interpreted the lack of diversity effects in terms of the Osherson et al. (1990) model, with the reasoning mechanism for generating and using a superordinate category not being well-developed in children.

Gutheil and Gelman (1997) made a further attempt to find evidence of diversity-based reasoning for specific conclusions in 9 year olds, using category members at lower, or more concrete, taxonomic levels which would presumably make reasoning easier, and also using increased sample size (e.g., three different butterflies rather than two different mammals). However, like Lopez et al. (1992), Gutheil and Gelman did not find diversity effects in 9 year olds, although in a control condition with adults, there was clear evidence for diversity effects with the same stimuli.

We see at least two 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). Reasoning in children might not be able to access all the same processes as reasoning by adults. 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.

Our own experiments were an attempt to distinguish between these two explanations, that the non-existence of diversity effects in children has been due to a lack of reasoning ability or due to lack of knowledge. We attempted to look for diversity-based reasoning in other domains, such as toys and clothing, that should be more familiar to children, compared to animals and their biological properties. If children show diversity-based reasoning for other kinds of categories, there would be evidence that the lack of diversity effects in past studies was due to incomplete knowledge in children rather than differing mechanisms for children's reasoning compared to adult reasoning.

Our procedure was similar to that used by Lopez et al. (1992) and Gutheil and Gelman (1997). The child was given a fact about one set of items, then another fact about another set of items. Then the child was asked which fact was true of a target item. To be more specific, 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 non-diverse choice). Our stimulus design was exactly analogous to past studies of diversity, using everyday objects and properties based on interactions with people. We tested children over a range of ages from 5 to 8 years, with the aim of looking for some evidence of diversity-based reasoning below age 9. Also, in the first experiment, we used two types of instructions, following Lopez et al. 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, England. The experiment was conducted on individual students; each session typically lasted 10 - 15 min.

Materials. There were 8 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 3 non-diverse items and a set of 3 diverse items. Each set was associated with a person. For example, in a non-diverse set there were three photographs of a football (soccer ball), being played with by a boy name 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 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 cm by 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 non-diverse 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 similarities between pairs of items.

Procedure. The order of test questions was randomized for each subject, and likewise on half the questions the non-diverse 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, then four test questions were given with emphatic instructions. The standard instructions involved presenting the 3 non-diverse photographs and 3 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 non-diversity or diversity. For example, the experimenter would say "Look, there's Tim. He's playing with the same thing, another football" for the non-diverse 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 4 test questions, we wanted to be certain that the diversity or non-diversity of each set was highly salient for each subject.

After the 6 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 feedback after the child's response. Otherwise, the experimenter never gave any reason for the child to favor either the non-diverse set or the diverse set in making predictions, in the standard instructions as well as the emphatic instructions.

Table 1. Stimuli for Experiment 1.

Target Item	Non-Diverse 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 are simplified. For example, the books varied in size and shape as well as color.

Table 2. Proportion of Diverse Choices, Experiment 1.

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

Results and Discussion

Overall, as shown in Table 2, children robustly favored the diverse choice over the non-diverse choice. For example, in the standard condition, the overall 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 ANOVA indicated a significant effect of instructions, $F(1,60)=8.66$, $p<.01$. The effect of year was not quite statistically significant, $F(3,60)=2.19$, $p<.10$, and the interaction was not close to the level of significance, $F(3,60)=1.31$. 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 8 stimulus sets, with mean proportion of diverse choices ranging from .69 to .90 and no significant item differences found. The 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. Indeed, we found diversity effects for flowers and pet animals, using social properties (human interaction or possession) rather than biological properties. (See Heit and Rubinstein, 1994, for further evidence on effects of properties.)

These results represent the first strong evidence for diversity-based reasoning in children under age 9. Indeed, we did not find major age differences in the range of 5 to 8 years. On the first four test questions, which did not emphasize the diversity or non-diversity of given items, children made the diverse choice 74% of the time overall. The proportion was even higher with emphatic instructions that highlighted diversity and non-diversity (but did not indicate which one to choose). This apparent effect of instructions, however, could also be a practice effect because the emphatic instructions were always for the last 4 items. (We could not present the emphatic instructions for the first 4 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 when they *don't* show diversity effects. Having a diverse set of premise categories should license a broad set of inferences compared to a non-diverse 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 non-diverse 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 weakened or even eliminated for the more superordinate target items.

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 pilot version of this experiment on 12 adults, comparing responses for 4 basic-level matches (e.g., another kind of doll) and 4 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 the oldest children in Experiment 1. For the superordinate-level items, the proportion of diverse choices was significantly lower, 69%, $t(11)=2.93$, $p<.05$.

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

Method

Experiment 2 was like Experiment 1, with the following changes. Ninety-two children, who attended Brookhurst Primary School, 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. For example, 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 non-diverse 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-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 were concerned about possible carry-over effects in which a student 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 will focus on these responses in this report.

Experiment 2 used the standard form of instructions from Experiment 1.

Results and Discussion

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, $F(1,88)=19.35$, $p<.001$, and school year, $F(1,88)=7.41$, $p<.01$. The interaction between these two variables did not approach statistical significance, $F<1$.

These results replicate and extend those of Experiment 1. The overall proportion of diverse choices for basic-level targets, 77%, is similar to that of the standard condition of Experiment 1, 74%. However, children were less likely to make the diverse choice for superordinate-level targets, 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 pilot study by adult subjects. Finally, the results of Experiment 2 supported an age effect, which was also suggested by Experiment 1. However, we would hesitate to over-interpret 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. 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.

In both experiments, the children sometimes gave explanations for their choices, and these explanations were recorded when possible. In Experiment 1, it was very obvious from talking with the children that they knew that one set was more diverse than the other, and that they were making inferences on this basis. For Experiment 2, the explanations for the superordinate-level items were much more idiosyncratic, suggesting that children were employing a variety of strategies, such as selecting items based on similarity within a single dimension such as size or color.

Table 4. Proportion of Diverse Choices, Experiment 2.

Year	Superordinate (S.E.)	Basic (S.E.)
1	.46 (.05)	.71 (.05)
4	.62 (.07)	.83 (.04)

General Discussion

In contrast to past studies, our results show clearly that children from age 5 to age 8 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 weakened diversity effects for remote inferences about more distantly related target categories. Therefore we would conclude that in terms of diversity, children can assess evidence and reason about categories in a manner similar to adults, if 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 are familiar enough. The past studies may have used materials that were too difficult or unfamiliar for children to support diversity-based reasoning. For example, Lopez et al. (1992) used properties such as "has leukocytes inside" and "has cellulose inside," rather than simple properties relating to associations with people.

We see this general issue to be important because there is a strong case to be made for the normative status of diversity-based reasoning. The value of diverse evidence for testing a hypothesis has been stressed repeatedly in the philosophical literature on scientific reasoning (e.g., Carnap, 1950; Nagel, 1939; Hempel, 1966). In brief, diversity-based reasoning is related to a falsifying test strategy; testing similar items repeatedly would be seen as a weak, confirmatory strategy. (See also Lopez, 1993.) Likewise diversity-based reasoning can be shown to be compatible with a Bayesian perspective (Heit, 1998; Howson and Urbach, 1993). Having such a powerful tool available early on in development thus seems of great adaptive value.

Acknowledgments

We are grateful to Jane Pollock for assistance in conducting this research. We thank the students and teachers of St Peter's Primary School and Brookhurst Primary School for their participation. This research was supported by a grant from BBSRC.

References

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.

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 University Press.

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.

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

Lopez, A. (1993). 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.

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

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

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

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