THINKING & REASONING, 2011, 17 (2), 156–181

Psychology Press
Taylor & Francis Group

Properties of the diversity effect in category-based inductive reasoning

Aidan Feeney[1] and Evan Heit[2]

[1]Queen's University Belfast, Belfast, Northern Ireland, UK
[2]University of California, Merced, CA, USA

Four experiments investigated how people judge the plausibility of category-based arguments, focusing on the diversity effect, in which arguments with diverse premise categories are considered particularly strong. In Experiment 1 we show that priming people as to the nature of the blank property determines whether sensitivity to diversity is observed. In Experiment 2 we find that people's hypotheses about the nature of the blank property predict judgements of argument strength. In Experiment 3 we examine the effect of our priming methodology on people's tendency to bring knowledge about causality or similarity to bear when evaluating arguments, and in Experiment 4 we show that whether people's hypotheses about the nature of the blank property were causal predicted ratings of argument strength. Together these results suggest that diversity effects occur because diverse premises lead people to bring general features of the premise categories to mind. Although our findings are broadly consistent with Bayesian and Relevance-based approaches to category-based inductive reasoning, neither approach captures all of our findings.

Keywords: Induction; Diversity; Relevance; Bayesian; Causality.

Correspondence should be addressed to Aidan Feeney, School of Psychology, Queen's University Belfast, University Road, Belfast BT7 1NN, Northern Ireland, UK. E-mail: a.feeney@qub.ac.uk

The authors gratefully acknowledge the very helpful comments made by Jean Baptiste van der Henst on an earlier version of this article. We thank Alex Whalley for statistical advice. We are also grateful to Robbie Gardiner, Michael Ewbank, Talin Tokatlian, Mary-Elaine McCavert, and Kathy Ruddy for their assistance with data collection and coding.

http://www.psypress.com/tar DOI: 10.1080/13546783.2011.566703

When we reason inductively we must decide how well some evidence supports a conclusion. For example, the fact that horses and cows possess some property P is evidence bearing on the conclusion "all mammals possess property P". One factor that determines the degree to which a conclusion is supported by some evidence is the diversity of that evidence. Philosophers have long argued that, all things being equal, a scientific theory is better supported by a set of diverse, rather than non-diverse, observations (Bacon, 1620/1898; for a review of the philosophical literature see Heit, Hahn, & Feeney, 2005). Although there has been some debate about the normative status (see Lo, Sides, Rozelle, & Osherson, 2002; Wayne, 1995), and universality (see Lopez, Atran, Coley, Medin, & Smith, 1997) of sensitivity to evidential diversity, such sensitivity has been found to emerge early in childhood (Heit & Hahn, 2001), particularly in pedagogical contexts (Rhodes, Gelman, & Brickman, 2010), to play a role in conceptual change during development (Hayes, Goodhew, Heit, & Gillan, 2003), to influence memory for learnt material (Hahn, Bailey, & Elvin, 2005), to play a role in hypothesis testing (Lopez, 1993; Rhodes, Brickman, & Gelman, 2008), to be important in diagnostic reasoning (Kim & Keil, 2003), and to be associated with cognitive ability (Feeney, 2007). Sensitivity to diversity is a widespread and useful aspect of cognition.

The effects of evidential diversity have been studied most often in the context of people's reasoning about categories. Osherson, Smith, Wilkie, Lopez, and Shafir (1990) presented participants with problems of the kind:

Horses have an ulnar artery. (1)
Seals have an ulnar artery.

All mammals have an ulnar artery.

Horses have an ulnar artery. (2)
Cows have an ulnar artery.

All mammals have an ulnar artery.

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. The features that people are asked to reason about in arguments of this sort are usually referred to as "blank predicates". In some cases, as in the example with which we began this paper, the predicate is referred to as "property X". In others such as arguments 1 and 2, predicates about which the participants are unlikely to possess knowledge are used. Unfamiliar or blank predicates are employed so that the effects of participants' knowledge about the categories in the argument may be studied independently of the effects of features.

With some exceptions (see next section) participants find problems specifying diverse evidence such as (1) above to be stronger than problems specifying non-diverse evidence such as (2), and early accounts of category-based diversity effects were based on the similarity between the categories in the premises and the category in the conclusions (see Osherson et al., 1990; Sloman, 1993). Our central aim in this paper is to describe a set of experiments on diversity effects which employed methods that are relatively novel for research on inductive reasoning, including priming manipulations as well as free response measures in which participants state their hypotheses about the nature of blank predicates. We will be most concerned with the implications of our results for recent Relevance and Bayesian accounts of the diversity effect, which we will briefly summarise below.

THE RELEVANCE-BASED APPROACH TO INDUCTION

The Relevance framework for inductive reasoning (Medin, Coley, Storms, & Hayes, 2003) is an attempt to capture effects that are problematic for similarity-based (Osherson et al., 1990; Sloman, 1993) accounts of category-based inductive reasoning. Those earlier accounts of induction are unable to explain how knowledge that is not about taxonomic or perceptual similarity might be involved in category-based inductive reasoning. According to Relevance Theory (Sperber & Wilson, 1986/1995), the relevance of some information is determined by its cognitive effects and by the cognitive effort required to process it, where relevance is increased by greater cognitive effects and decreased by more cognitive effort. In the context of a category-based inductive argument, cognitive effects might include the updating of beliefs about the conclusion or the generation of hypotheses about the blank predicate. Cognitive effort, on the other hand, is determined by the extent to which knowledge about relations between the categories in the argument is easily available.

In applying Relevance Theory to category-based induction Medin et al. emphasised the importance of people's beliefs about the blank predicate and suggested two principles that guide people's evaluation of category based inductive arguments. The first is that people search for the most distinctive features of the categories in the premises. Thus, when presented with the single premise "All cheetahs have property A" people may assume that, in order to identify the blank property, the experimenter expects them to think about speed, a distinctive and highly available property of cheetahs. The second principle is that people compare the categories in the premises to the category in the conclusion as a way of testing the projectability of candidate features. For example, if the premise concerns cheetahs and the conclusion is that "greyhounds possess Property A" then, as speed is also a salient property of greyhounds, expectations of relevance raised by the premise regarding the feature "speed" will be confirmed by the conclusion and the reasoner will conclude that speed is likely to be the relevant property and that the argument is a strong one.

Medin et al. (2003) demonstrated a number of non-diversity effects that suggest the importance of people's guesses about the to-be-projected property. For example, they presented participants with pairs of arguments such as 3 and 4 below:

Camels have Property X. (3)
Rhinos have Property X.

Mammals have Property X.

Camels have Property X. (4)
Desert rats have Property X.

Mammals have Property X.

The premises in Argument 4 concern diverse categories that share a highly available property (shared habitat) unlikely to generalise to the conclusion category. Argument 3 concerns less diverse categories that do not share a salient property. Medin et al. predicted a *non-diversity effect by property reinforcement*. That is, they expected Argument 4 to be weaker than Argument 3 because in the former case participants would suspect that the blank property was similar to the shared non-projectible feature. In order to demonstrate a second phenomenon that they called *causal non-diversity*, Medin et al. presented participants with pairs of arguments concerning diverse but causally related, or less diverse but non-related, premise categories, such as 5 and 6 below:

Koalas have Property X. (5)
Wolves have Property X.

Living Things have Property X.

Koalas have Property X. (6)
Gum Trees have Property X.

Living Things have Property X.

Although the premise categories in Argument 6 are more diverse than those in Argument 5, the researchers predicted that the causal relationship between koalas and gum trees would weaken the generalisation from those categories to the superordinate category *living things*. As they predicted, Medin et al. found significant non-diversity effects in both cases.

Although recent evidence supports the Relevance account of inductive reasoning generally (Coley & Vasilyeva, 2010; Feeney, Coley, & Crisp, 2010), experimental support for the Relevance account of diversity is mixed. Consistent with predictions from the Relevance framework, Feeney et al. found that property reinforcement resulted in a failure to observe diversity effects. However, even stronger evidence for the theory would have involved significant non-diversity effects. Heit and Feeney (2005) suggested that non-diversity effects due to property reinforcement can be explained in terms of similarity. Using Medin et al.'s materials, Heit and Feeney replicated the non-diversity by property reinforcement effect, but showed that people viewed the categories in the stronger argument as less similar than the categories in the weaker argument. For example, they considered penguins and polar bears to be more similar than polar bears and antelopes. Indeed, similarity ratings strongly predicted judgements of inductive strength. Thus, Medin et al.'s property reinforcement materials produce a diversity effect and may thus be explained by similarity-based models. This finding leaves the causal non-diversity effect as the sole distinctive piece of evidence concerning the diversity effect that has been produced in favour of the Relevance approach. For this reason, we studied causal non-diversity effects in two of the experiments to be reported here.

BAYESIAN ACCOUNT OF CATEGORY-BASED INDUCTION

Bayesian accounts of inductive reasoning (see Heit, 1998, 2000; Kemp & Tenenbaum, 2009) offer a computational-level (Marr, 1982) hypothesis-testing account of category-based induction. The hypotheses in these accounts concern the nature of the feature being reasoned about and are informed by background knowledge about how features are distributed. The model assigns prior beliefs to these hypotheses, which in the light of the information presented in the premises, it then revises using Bayes' Theorem.

To illustrate how such models explain reasoning, consider how they handle Argument 1 above. The models suggest that people's goal is to decide whether possession of an ulnar artery is likely to be an all-mammals property (H1), a horses-and-seal property (H2), a horse-only property (H3), a seal-only property (H4), or a property possessed by neither horses nor seals (H5). On the basis of background beliefs about feature distributions, the model assigns priors to these hypotheses. Using Bayes' Theorem (Equation 1) the models calculate how belief in each hypothesis should be revised upon receipt of the information in the premises.

$$P(H_i/D) = \frac{P(H_i)P(D/H)}{\sum_{j=1}^{n} P(H_j)P\left(D/H_j\right)} \qquad \text{(Equation 1)}$$

Bayesian models differ in how they assign priors. According to Heit (1998, 2000) these priors may be driven by the kinds of features that come to mind. Because horses and seals are dissimilar, the features that come to mind are very general. That is, most of the properties they share are likely to be shared by other mammals. Accordingly, H1 is assigned a high prior probability and H2 is assigned a relatively low probability. As horses and seals are dissimilar it is also likely that one has features not shared with the other, so H3 and H4 may be assigned quite high probabilities. It is also likely that no animals possess the feature so H5 may also be assigned a relatively high probability. If we consider Argument 2 above (Horses and cows share a property, therefore all mammals possess the property), background knowledge about feature distributions leads to the assignment of different priors to the hypotheses. As horses and cows are similar, they are likely to share many features and the features that come to mind are likely to be idiosyncratic to mammals like horses and cows. Accordingly, the hypothesis that all mammals possess the feature is assigned a relatively low probability and the hypothesis that only horses and cows possess the feature is assigned a relatively high probability. As horses and cows are similar, it is also less likely that one on its own will possess a feature than it was in the case of Argument 1, so H3 and H4 are assigned relatively low probabilities. These different prior probabilities lead to the model predicting a diversity effect for Arguments 1 and 2,

In Heit's model, when evaluating arguments containing blank predicates, priors for each hypothesis can be driven by the features that come to mind. An alternative approach to prior determination has been taken by Kemp and Tenenbaum (2009; see also Shafto, Kemp, Bonawitz, Coley, & Tenenbaum, 2008) who have described a class of models where priors are determined from people's intuitive theories about the domains.

These theories take the form of structured relations between categories and how features are distributed across those categories. Kemp and Tenenbaum considered several different kinds of structure including a default tree structure for the taxonomic domain and a web structure for taxonomic contexts that specifically concern food chains. In this approach, people's beliefs about the property determine which type of structured representation they bring to mind. When the property is blank then the default tree-structure is brought to mind. This approach accounts for diversity effects in the taxonomic domain by virtue of diverse premise sets making it more likely that the blank property has spread throughout the tree whereas non-diverse sets make it more likely that the property is localised to a particular branch of the tree. It also accounts for causal non-diversity effects by virtue of its ability to estimate priors over a web structure representing food chain relationships rather than over a taxonomic tree structure.

THE EXPERIMENTS

The four experiments in this paper used two novel methods to study diversity effects. Both methods were prompted by Heit's (1998, 2000) suggestion that non-diverse premise sets cause idiosyncratic, and therefore less projectible, properties to come to mind whereas general properties come to mind when evaluating arguments with diverse premises. Although an idiosyncratic property might be projected onto a conclusion category that is similar to the premise categories, a general property could be projected more broadly.

In Experiment 1 we employed a novel feature priming methodology where we asked people to evaluate arguments with diverse and non-diverse premise categories in the presence of general or idiosyncratic feature primes. Our central prediction here was that prime type and premise diversity should determine ratings of argument strength. In Experiment 2 we asked participants, in the absence of a prime, to record their hypotheses about the nature of the blank feature when evaluating diverse and non-diverse arguments.

We also used our priming and feature listing methods to explore the causal non-diversity effect. In particular we aimed to replicate the causal non-diversity effect (Experiments 3 and 4); to examine the effects of priming people about the existence of a causal relationship (Experiment 3); and to test whether decreased ratings of argument strength are associated with hypotheses that the blank property concerns a causal relation between the premise categories (Experiment 4).

EXPERIMENT 1: PRIMING INDUCTION

To test the prediction that priming participants with a general or an idiosyncratic feature shared by the premise categories of an inductive argument would affect judgements of argument strength, we presented participants with diverse and non-diverse arguments alongside idiosyncratic and general primes. These primes were presented as an example of the features shared by the categories in the premises. Our manipulation is more direct than those typically used in the study of semantic and associative priming (for a review, see Hutchison, 2003). Our own aim was not to establish whether induction can be primed or how sensitive induction is to priming, but rather to directly compare the effects of two different types of primes. In addition, because models of induction make predictions based on what information is available, it was useful to have an explicit manipulation of what is available.

We pre-tested prime features for frequency and to ensure that people thought that they were shared by the premise categories of the argument. We predicted a standard diversity effect such that diverse arguments should be rated stronger than non-diverse arguments. We also predicted a priming effect so that general primes should produce stronger ratings of argument strength than idiosyncratic primes.

Pre-testing feature frequency

We pre-tested features for use as primes in two different ways. Initially we asked a group of 20 students at Durham University to estimate the percentage of all mammals that possess a series of 31 features. The majority of these features were expected to be idiosyncratic and were initially chosen on the basis of our own intuitions. They were presented in one of two random orders.

Participants' frequency estimates for each of the features finally selected for inclusion in the induction part of the experiment are presented in Table 1. All idiosyncratic features chosen for inclusion in the second pre-test had a mean prevalence rate of less than 50%. The mean prevalence rate for the idiosyncratic features was 30% ($SD = 10.66$). The equivalent mean percentage for the general features was 94% ($SD = 7.21$). The difference between these means is highly reliable, $t(19) = 21.60$, $p < .001$.

Estimates of shared properties

The second pre-test was designed to confirm our intuitions about pairing idiosyncratic and general features with mammal pairs to be used in the induction task. We used the same mammal pairs for which Osherson et al. (1990) had obtained pair-wise similarity ratings. We presented a further 20 student participants from Durham University with each pair of mammals and asked them to estimate on a scale from 1–10 the extent to which they believed that the pair shared each of five features. One of these features was always a feature rated as general in the first pre-test, while the other four had been rated as idiosyncratic. Participants were asked about each pair of mammals in one of two randomly determined orders.

The mean ratings from this task for each of the features used in the induction task are to be found in Table 1. Our criterion for inclusion of a feature was that the mean rating be over 5. All features satisfied this condition except the idiosyncratic feature chosen for presentation with the pair Elephants and Squirrels. This feature was selected for inclusion as it received the highest mean rating of all the idiosyncratic features pre-tested for this pair.

Ratings of inductive strength

Participants. A total of 40 students and members of staff from Durham University participated in this experiment.

TABLE 1
Stimuli and results from Experiment 1

	Primes and probability judgements							
	General prime				*Idiosyncratic prime*			
Premise categories	*Feature*	*Feature frequency*	*Shared by categories*	*Inductive rating*	*Feature*	*Feature frequency*	*Shared by categories*	*Inductive rating*
Non-diverse pairs								
Dolphin–Seal	Heart	100.0%	9.0	53.4%	Live in water	14.1%	8.3	45.6%
Mouse–Squirrel	Teeth	93.9%	9.2	58.2%	Whiskers	41.0%	7.7	38.7%
Horse–Cow	Warm-blooded	91.0%	9.0	62.0%	Hooves	27.6%	8.7	30.2%
Elephant–Rhinoceros	Four limbs	86.5%	8.3	62.8%	Thick skin	41.1%	9.1	33.0%
				59.1%				36.9%
Diverse pairs								
Elephant–Squirrel	Tongue	93.8%	7.9	93.9%	Eat nuts	22.8%	4.3	37.3%
Horse–Seal	Internal skeleton	98.0%	9.3	91.0%	May be trained	39.0%	8.2	61.1%
Cow–Dolphin	Brain	97.0%	9.2	86.5%	Gentle	34.7%	5.2	35.9%
Rhinoceros–Mouse	Lungs	94.5%	9.2	100.0%	Stumpy legs	16.7%	5.4	42.3%
				68.0%				44.1%

Design. The experiment had a $2 \times 2 \times 2$ mixed design. Diversity and prime type were manipulated between participants and form of materials was manipulated between participants. Thus participants gave ratings of inductive strength to four diverse and four non-diverse arguments. In addition, half of each set of four arguments was presented alongside a general prime and half with an idiosyncratic prime. We counterbalanced the pairing of prime and argument resulting in two versions of the materials. Each of these versions was presented in two different orders or forms, form A and form B.

Materials

Participants evaluated eight inductive arguments concerning properties of animals labelled consecutively from Property A to Property H. For each question the participant was given two facts to assume to be true about particular kinds of mammals. Then the participant judged the probability that this fact would apply to "all mammals". Their instructions were to make the best possible estimate based on general knowledge about biological properties of animals.

Each argument was presented alongside a prime. The prime consisted of a reminder that the mammals mentioned in the premises shared many features and gave an example of a general or an idiosyncratic feature. An example of an argument presented with a general prime is given below:

Horses and cows share many features. For example, they are both warm-blooded.
Given the fact that:
Horses have property A.
Cows have property A.

How likely is it that:
All mammals have property A?

Response: (0% to 100%) ____%

To counterbalance the pairing of primes with arguments, we split the eight arguments into randomly determined halves. The single constraint was that each half contained two diverse and two non-diverse pairs. Two versions of the problem materials were assembled. These versions differed in terms of whether the problems in each half were presented with a general or an idiosyncratic prime. Thus each of the arguments was rated by half of our participants in the presence of a general prime and by the other half in the presence of an idiosyncratic prime. Both counterbalanced versions of the materials were presented in one of two forms, which differed in the order of the questions. Half the participants were given form A of the questionnaire and half were given form B.

Results and discussion

The mean probability ratings for each argument, broken down by prime type, are presented in Table 1. The mean rating of inductive strength for the diverse arguments (56.1%) was greater than the equivalent mean for non-diverse arguments (48.0%). In addition, mean ratings of inductive strength were higher for arguments presented with a general prime (63.6%) than with an idiosyncratic prime (40.5%).

The results of a 2 (diversity) × 2 (prime type) × 2 (form) mixed design ANOVA confirmed these observations. There was a significant effect of diversity, $F(1, 38) = 4.43$, $MSe = 591.08$, $p < .05$. There was also a significant main effect of prime, $F(1, 38) = 33.04$, $MSe = 643.40$, $p < .001$. Although the effects of diversity and prime are clearly additive rather than interactive (see Table 1), it is interesting to observe that mean ratings of inductive strength for non-diverse arguments presented with a general prime are greater than mean ratings for diverse arguments presented with an idiosyncratic prime. Finally, although the effect of form was not significant, $F(1, 38) < 1$, the three-way interaction between form, argument type, and prime type was significant, $F(1, 38) = 7.05$, $MSe = 71.30$, $p < .05$. However, the key pattern of results did appear regardless of the order in which the questions were presented.

The results of this experiment have replicated the diversity effect and confirmed our prediction that general primes lead to higher ratings of argument strength than idiosyncratic primes. Although this prediction was directly derived from Heit's (1998) Bayesian account of the diversity effect, as we will see in the General Discussion it can also be explained by the other accounts that we have discussed.

EXPERIMENT 2: FEATURE LISTING AND INDUCTIVE STRENGTH

In Experiment 2 we tested more directly the assumption that general features are more likely to come to mind when people reason about diverse categories and that idiosyncratic features will come to mind when the premises concern non-diverse categories. To obtain more direct evidence for this claim, in Experiment 2 we asked participants to tell us their best hypothesis about the nature of the blank property in each argument (see Blok & Gentner, 2000, for a related technique). This method allowed us to test a second prediction: that ratings of argument strength should be associated with the rarity of the feature that comes to mind. That is, regardless of whether an argument is diverse or non-diverse, if during the evaluation of an argument an idiosyncratic feature comes to mind then that argument should receive lower strength ratings than when a general feature comes to mind.

Method

Participants. A total of 40 University of Warwick students participated for payment of one pound.

Materials. There were two paper-based questionnaires, for the probability judgement task and for the property-listing task. The probability judgement task was similar to the inductive task used in Experiment 1 except that primes were not presented alongside the arguments. The other difference between it and the task used in Experiment 1 was that it had 11 questions, 3 of which were filler questions about just a single given fact. The remaining eight critical questions were identical to those used in Experiment 1.

The property-listing task had the same 11 items as the probability judgement task, in the same order. Participants were asked what they thought each of the properties really was, and to answer in a few words or a brief sentence. They were reminded that the property should be true of the animal or animals in the premise(s) of the argument. Participants were encouraged to think back to how they originally judged the probability of each statement.

Half the participants were given form A of both questionnaires and half were given form B. The two forms differed in the order of the questions. Also, the forms differed in the order of the facts listed within each question. For example, in form B the facts were given about squirrels and elephants rather than elephants and squirrels.

Procedure. Participants were tested individually in a quiet place. Each participant completed the probability judgement task, then this questionnaire was taken away. The probability judgement task was followed immediately by the property-listing task.

Results

Probability judgements. The mean probability judgements for the eight critical stimuli are shown in Table 2. The average judgement for diverse pairs, 67.1%, is greater than the average judgement for non-diverse pairs, 44.5%, and indeed all of the diverse pairs received higher judgements than any of the non-diverse pairs. This result suggests that the overall diversity effect was replicated. A two-way analysis of variance (ANOVA) supported this observation. The effect of diversity was statistically significant, $F(1, 38) = 61.89$, $MSe = 0.020$, $p < .001$. The effect of form was not statistically

TABLE 2
Stimuli and results from Experiment 2

Premise categories	*Probability judgement*	*Proportion of idiosyncratic properties*
Non-diverse pairs		
Dolphin–Seal	39.3%	72.5%
Mouse–Squirrel	44.9%	55.0%
Horse–Cow	46.9%	57.5%
Elephant–Rhinoceros	51.6%	72.5%
	44.5%	64.4%
Diverse pairs		
Rhinoceros–Mouse	67.7%	22.5%
Elephant–Squirrel	68.4%	35.0%
Horse–Seal	72.4%	22.5%
Cow–Dolphin	73.8%	15.0%
	67.1%	23.7%

significant, $F(1, 38) = 0.023$, $MSe = 5.19$, and the interaction between diversity and form was not significant, $F(1, 38) = 2.22$, $MSe = 0.020$.

Property listings. The participants seemed to have understood the property-listing task and generally wrote relevant properties such as "has eyes", "is warm blooded", "is large", and "lives in the water". The 320 responses were coded independently by two research assistants who were not informed of the experimental hypotheses. The properties were coded into two categories, idiosyncratic (applies to few or some mammals) and general (applies to most or all mammals). For example, "has eyes" and "is warm-blooded" are general properties whereas "is large" and "lives in the water" are idiosyncratic properties. The two coders agreed initially on 95.6% of the responses. The remaining conflicts were resolved by discussion and, when necessary, consultation of reference books.

The mean proportions of idiosyncratic properties listed for the eight stimuli are shown in Table 2. The average proportion for non-diverse pairs, 64.4%, is greater than the average proportion for diverse pairs, 23.7%, and likewise all of the non-diverse pairs had higher proportions of idiosyncratic properties than any of the diverse pairs. This finding supports the hypothesis that the diversity effect would be associated with generating properties that apply to most mammals, whereas for non-diverse pairs people would tend to generate idiosyncratic properties. An ANOVA corroborated this observation. The effect of diversity was statistically significant, $F(1, 38) = 46.48$, $MSe = 0.0071$, $p < .001$. The effect of form was not statistically significant,

$F(1,\ 38) = 0.01$, $MSe = 0.0074$, and the interaction between diversity and form was not significant, $F(1,\ 38) = 1.33$, $MSe = 0.0071$.

The preceding ANOVA compares the non-diverse items as a set to the diverse items as a set. In addition, with the eight stimuli treated individually, as predicted by Bayesian and Relevance accounts, there was a strong association between probability judgements and proportion of idiosyncratic properties generated. The correlation across the eight stimuli, in effect between the second and third columns of Table 2, is very strong: -0.923, $p < .001$.

Although the above analyses show a strong relation between the probability judgements and the property listings, they are taken at the group level and thus do not assess performance for individual participants. However, the properties listed by each individual participant were in fact good predictors of that participant's probability judgements. Of the 40 participants, all of one person's properties were coded as idiosyncratic and all of another person's properties were coded as general. The remaining 38 participants gave a mix of the two kinds of properties. For each of these participants the average probability judgement for items leading to an idiosyncratic property was calculated, and likewise the average judgement for general properties was calculated. Taken across participants, the average judgement for idiosyncratic properties was 52.8% and the average judgement for general properties was 65.1%. This difference was statistically significant by a paired t test, $t(37) = 3.59$, $p < .001$. Hence, at an individual level, generating an idiosyncratic property was associated with lower probability judgements than generating a general property.

Discussion

The results of Experiment 2 are consistent with the idea that feature idiosyncrasy plays an important role in the diversity effect. Participants were almost three times more likely to list an idiosyncratic feature for non-diverse premise sets than they were for diverse premise sets. In addition, the proportion of idiosyncratic features generated was a good predictor of argument strength, a finding that held across items and participants. When considered in conjunction with the results of Experiment 1, these results suggest that diversity effects are observed because diverse premises cause people to bring to mind general features shared by the premise categories, whereas non-diverse premises cause people to bring idiosyncratic and hence less projectible features to mind.

EXPERIMENT 3: FEATURE PRIMING AND CAUSAL NON-DIVERSITY

In the third experiment in this series we employed the priming methodology used in Experiment 1 to examine causal non-diversity. This effect is of particular interest because some other non-diversity effects originally offered as support for the Relevance framework can be explained on the basis of similarity (see Heit & Feeney, 2005).

In this experiment, using the materials originally devised by Medin et al. to demonstrate causal non-diversity, we contrasted the case where diverse and non-diverse arguments are primed with a general feature to the case where they are primed with a causal relationship. Medin et al. compared two types of argument. Arguments of the first type were based on diverse premises that made assumptions about the causal relationship between their categories particularly accessible (e.g., going from koalas and eucalyptus trees to all living things). The premises of the second type of argument were less diverse but did not make assumptions about a causal relationship between their categories available (e.g., going from koalas and wolves to all living things). Medin et al. observed a significant non-diversity effect such that arguments of the first type were judged to be weaker than arguments of the second type. The first aim of this experiment was to examine whether, collapsed across prime type we could replicate the causal non-diversity effect. The second aim was to test for an effect of our priming manipulation such that arguments presented with a general prime should be significantly stronger than arguments presented with a causal prime.

Method

Participants. A total of 58 students of psychology at the University of Durham participated in this experiment.

Materials. The experiment had a $2 \times 2 \times 2$ mixed design. Causal relation between the premise categories and prime-type were manipulated between participants. Each participant provided ratings of inductive strength for 12 arguments. Six of these arguments concerned causally related categories and six concerned categories that were unrelated. The premise categories in the arguments were taken from Medin et al. (2003). Whereas Medin et al. used specific and general conclusion categories for different groups of participants, the conclusion category *all living things* was used for the arguments presented to participants in this experiment. The general primes used in this experiment were taken from Experiment 1. The causal primes were causal relationships between the categories. For the causally related pairs these relationships were the salient relationship. For instance, in an argument where koalas and eucalyptus trees were the premise categories, participants read "*Koalas and eucalyptus trees have many things in common. For example, koalas live on eucalyptus trees*". For the non-causally related categories we invented causal relationships that were modelled on the relationships between the causally related category pairs. All premise pairs and primes are presented in Table 3.

We randomly divided the categories into two sets with the constraint that each set should contain three causally related and three non-related problems. In half of the experimental booklets one set was presented in the presence of a general prime and the other set in the presence of a causal prime. This pairing of set and prime was reversed for the remaining booklets. Thus each participant rated six arguments in the presence of a general prime and six in the presence of a causal prime. However, they rated each argument only once so, for each argument, prime was manipulated between participants. Half the participants were given form A of the questionnaire and half were given form B. The two forms differed in the order of the questions. The order of the questions in form A was the same for both sets of problems as it was in form B.

Participants responded on each problem by giving a percentage to indicate how strong they thought the argument to be. As was the case in Experiment 1, the biological features of the categories were emphasised in the instructions.

TABLE 3
Stimuli and results from Experiment 3

	Primes and probability judgements			
	General prime		*Causal prime*	
Premise categories	*Feature*	*Rating*		*Rating*
Causal diverse pairs				
Cats and Sparrows	Contain RNA	59.4%	Cats catch sparrows	40.7%
Fleas and Dogs	Mitochondria	60.9%	Fleas live on dogs	44.9%
Horses and Grass	Contain salts	57.2%	Horses eat grass	39.2%
Rabbits and Carrots	Contain DNA	66.1%	Rabbits eat carrots	46.6%
Sparrows and Seeds	Contain protein	60.6%	Sparrows eat seeds	34.9%
Koalas and Eucalyptus trees	Contain water	57.4%	Koalas live on eucalyptus trees	43.5%
		60.3%		41.6%
Non-causal less-diverse pairs				
Cats and Rhinos	Contain RNA	59.2%	Live in close proximity	39.7%
Fleas and Butterflies	Mitochondria	52.3%	Fleas eat butterfly larvae	43.4%
Sparrows and Dogs	Contain water	56.7%	Dogs prey on sparrows	41.6%
Koalas and Wolves	Contain water	68.3%	Wolves prey on koalas	42.7%
Horse and Ants	Contain salts	53.9%	Ants eat horse droppings	41.6%
Rabbits and Zebras	Contain DNA	66.7%	Eat grass	42.0%
		59.5%		41.8%

Results and discussion

The mean rating of inductive strength for each argument, broken down by prime type, is presented in Table 3. The overall rating for arguments presented with a general prime (59.9%) is higher than for arguments presented with a causal prime (41.7%). Thus we replicated the priming effect observed in Experiment 1. However, the mean rating given to the causally related arguments (51.1%) is about the same as the mean for non-causally related arguments (50.5%).

The results of a $2 \times 2 \times 2$ mixed design ANOVA confirmed that the main effect of prime-type was highly significant, $F(1, 56) = 29.46$, $MSe = 648.26$, $p < .001$. None of the other effects tested by the ANOVA was significant (all Fs < 1).

The results of this experiment provide novel evidence for the role of causal knowledge in inductive reasoning, in this case knowledge of food chains and proximity of habitat. Priming the existence of a causal relationship between the categories in the premises of an inductive argument has a marked effect on ratings of argument strength. In particular, the diversity effect was weakened when participants were reminded of a causal relation between a pair of diverse categories. However, we did not succeed in replicating the causal non-diversity effect and so one aim of the final experiment in this series was to replicate Medin et al.'s original effect.

EXPERIMENT 4

In Experiment 4 we employed the feature listing method from Experiment 2 to further examine Medin et al.'s causal non-diversity effect. We had two aims in this experiment. First we attempted to replicate the original effect. We did not observe an overall non-diversity effect in Experiment 3, but because we paired primes with arguments in the experiment, it may not have been a fair test of Medin et al.'s original hypothesis. Hence in the first part of Experiment 4 participants rated arguments designed to test for non-diversity in the absence of primes. Our second aim was to test the role of hypotheses about the nature of the blank property in ratings of argument strength. Specifically, we wished to test the hypothesis, derived from the Relevance approach, that participants who hypothesised that the blank property was in some way connected to a causal relationship between the premise categories would produce weaker ratings of argument strength than participants who did not hypothesise on the basis of a causal relationship. Recent work by Coley and Vasilyeva (2010) strengthens our hypothesis. Those authors showed that salient causal relations between the categories in an argument lead participants to spontaneously generate causal inductive inferences. In the case of arguments with multiple premises, inferring that the property is connected to a causal relationship between the premise categories should weaken argument strength.

Method

Participants. A total of 45 students from the University of Durham participated in the experiment.

Materials. There were two paper-based questionnaires, for the probability judgement task and for the property-listing task. The probability judgement task was similar to the inductive task used in Experiment 3 except that primes were not presented alongside the arguments.

The property-listing task had the same 12 items as the probability judgement task, in the same order. Participants were asked what they thought each of the properties really was, and to answer in a few words or a brief sentence. They were reminded that the property should be true of the animal or animals in the premise(s) of the argument. Participants were encouraged to think back to how they originally judged the probability of each statement.

Half the participants were given form A of both questionnaires and half were given form B. The two forms differed in the order of the questions.

Procedure. Participants were tested at the beginning or end of lectures. Each participant completed a booklet containing the probability judgement task and the property-listing task in that order.

Results

Probability judgements. The mean probability judgements for the 12 critical stimuli are shown in Table 4. The average judgement for the causally related arguments, 45.5%, is close to the average judgement for the non-causally related arguments, 47.3%, and in only four out of six items was the causally related argument rated weaker than the non-causally related argument. This result suggests that the overall causal non-diversity effect from Medin et al. (2003) has not been replicated. A two-way analysis of variance (ANOVA) supported this observation. The effect of argument type did not approach statistical significance, $F(1, 43) = 0.30$, $MSe = 234.71$, $p = .59$. The effect of form was not statistically significant, $F(1, 43) = 0.58$, $MSe = 492.79$, $p = .45$, and the interaction between diversity and form was not significant, $F(1, 43) = 1.17$, $MSe = 234.71$, $p = .29$.

Property listings. Not all participants listed a property for every argument (including one participant who had listed no features). The 496

TABLE 4
Stimuli and results from Experiment 4

Premise categories	*Probability judgement*	*Proportion of causal properties*
Causal diverse pairs		
Koalas and Eucalyptus trees	44.8%	51.2%
Horses and Grass	48.6%	52.3%
Rabbits and Carrots	43.5%	67.6%
Fleas and Dogs	43.4%	26.8%
Cats and Sparrows	48.1%	12.2%
Sparrows and Seeds	44.2%	42.1%
	45.5%	41.6%
Non-causal less-diverse pairs		
Koalas and Wolves	46.8%	2.4%
Horse and Ants	48.8%	9.3%
Rabbits and Zebras	49.7%	0.0%
Fleas and Butterflies	40.2%	0.0%
Cats and Rhinos	47.8%	0.0%
Sparrows and Dogs	50.5%	7.5%
	47.3%	2.7%

(out of a possible 528) responses were coded independently by two research assistants who were not informed of the experimental hypotheses. The properties were coded into three categories: causal transmission, shared habitat, and other. In the instructions to coders, causal properties were defined as follows: "Responses which suggest that the property has been transmitted from members of one category to members of the other. This might be because the response explicitly mentions transmission, or that members of one category tend to prey on members of the other, or does not mention transmission but mentions a property that could be transmitted from one to the other." Shared habitat responses were defined as those suggesting that the property is shared by category members because they share a habitat. The two coders agreed initially on 97.7% of the responses. Half of the remaining conflicts were resolved by discussion between the coders, and the remainder by consultation with one of the authors. Causal transmission properties accounted for 17.3% of all responses and shared habitat properties for 4.6% of responses. For the purposes of statistical analysis we combined causal transmission and shared habitat features into a more general causal category on the grounds that shared habitat properties are consistent with a common cause model (e.g., see Sloman, 2005) of how members of both premise categories came to possess the property.

Calculated across all participants who listed a property for that argument, the mean proportions of causal properties listed for the 12 stimuli are shown in Table 4. The average proportion of causal properties listed for the causal diverse pairs, 42%, is greater than the average proportion for the remaining non-causal non-diverse pairs, 3%, and likewise all of the causal pairs had higher proportions of causal properties than any of the non-causal pairs. An ANOVA corroborated this observation. The effect of argument type was statistically significant, $F(1, 42) = 60.88$, $MSe = 0.06$, $p < .001$. The effect of form was not statistically significant, $F(1, 42) = 0.18$, $MSe = 0.06$, and the interaction between diversity and form was not significant, $F(1, 42) = 0.09$, $MSe = 0.06$.

As in Experiment 2, we examined the correlation between probability judgements and feature listings across items. Because the non-causal pairs had few causal features listed, we focused on the causal pairs. Across the six items, the correlation between probability judgements and proportion of causal properties, corresponding to the top half of Table 4, was $-.31$, which did not reach the level of statistical of significance. This is admittedly a low-powered analysis, but the sign of the correlation was consistent with the idea that causal properties were associated with a weakening of the diversity effect.

To provide a stronger analysis of the relation between the probability judgements and the property listings, we tested whether individual participants' probability judgements were lower when they listed causal properties than when they listed non-causal properties. To do so we focused on the causally diverse pairs, because the non-diverse pairs had so few causal properties listed, and we only used responses from the 36 participants without any missing data (recall that some participants omitted one or more property listings). The question of interest was simply whether probability judgements associated with causal properties were lower than probability judgements associated with non-causal properties. In fact, the mean probability judgement when causal properties were listed was 36.8%, and the mean probability judgement when non-causal properties were listed was 53.6% (this pattern held when we included all participants and all arguments, regardless of missing data). Because each participant contributed probability judgements for six items to the analysis, and because several participants only listed either causal or non-causal properties, we used a version of the generalised linear model that accommodates clustered data (as implemented in the generalised estimating equation module in SPSS Version 16). There was a statistically significant effect of type of property listed (causal or non-causal), Wald's chi-squared $= 12.16$, $p < .001$, and neither the effect of item or the interaction between type of property and item were statistically significant.

Discussion

This experiment produced two main findings. First, we were unable to replicate the causal non-diversity effect, and in the next section we will discuss the implications of this failed replication for the Relevance framework. However, we have shown that more causal features are listed for causal arguments and that those causal arguments for which participants listed a causal feature were rated weaker than were causal arguments for which participants did not list a causal feature. As well as having implications for theories of category-based induction, this latter finding supports more general claims about the role of causal knowledge in inductive reasoning (see Rehder, 2007).

GENERAL DISCUSSION

The experiments described in this paper add to our understanding of the diversity effect in category-based induction, and suggest that diversity effects are due to diverse premises causing people to bring general features to mind. In support of this claim we have shown that priming participants with idiosyncratic features (Experiment 1) or a causal relationship (Experiment 3) significantly decreases argument strength. In addition, the generation of an idiosyncratic feature (Experiment 2) or causal relationship (Experiment 4) is associated with weaker ratings of argument strength. Another important finding from these experiments is that although we replicated the basic diversity effect in Experiments 1 and 2, we were unable to replicate the causal non-diversity effect in two experiments (Experiments 3 and 4). We will now consider the implications of these findings for the Relevance framework and Bayesian accounts of inductive reasoning.

Implications for the Relevance framework

The priming effects observed in Experiments 1 and 3 can easily be accommodated under the Relevance framework for inductive reasoning. Relevance Theory (Sperber & Wilson, 1995, pp. 38–46) predicts that many assumptions may be made available in a particular processing context. For example, the following argument taken from Medin et al. (2003) makes several assumptions manifest.

Robins have Property X. (5)
Worms have Property X.

Goldfish have Property X

One assumption likely to be highly available is that the experimenter intends to make manifest that the causal relationship that exists between robins and worms is relevant to the nature of the blank predicate. Another, perhaps less-available, assumption might be that the experimenter intends to make manifest that the dissimilarity between robins and worms is relevant. If the first assumption is most available in a particular context, the argument will be judged weak because goldfish are not part of the same causal relationship whereas, if the second assumption is most available, the argument will be judged strong, because the property applies to a diverse range of organisms.

One way to view the priming manipulation employed in Experiments 1 and 3 is that, in relevance-theoretic terms, it is ostensive. That is, participants interpret the primes as indicating an intention on the part of the experimenter to make something manifest to them. In this way a particular prime may affect the availability of various contextual assumptions. For example, the presentation of a general prime may make assumptions that the experimenter's communicative intentions related to the dissimilarity between the categories particularly accessible. An idiosyncratic prime, on the other hand, may make available assumptions that the experimenter intended to convey that similarities between the premise categories is relevant to the nature of the blank predicate. Thus the priming findings from Experiments 1 and 3 can be accommodated under the Relevance framework.

The results of our feature-listing experiments (Experiments 2 and 4) may also be explained by the Relevance view, which is that participants generate hypotheses about the nature of the blank predicate. In the case of the diversity effect replicated in Experiment 2 where diverse premise sets make available the assumption that the experimenter intends to convey the dissimilarities between the premise categories, those hypotheses are likely to concern very general features shared by even dissimilar animals. The results of Experiment 4, where we showed that arguments for which causal features were listed were rated weaker than arguments for which non-causal features were listed, also support the Relevance framework. Presumably those arguments for which causal features are listed make available assumptions that the experimenter intended to convey that the causal relation between the premise categories is relevant.

Although our priming and feature listing effects can be explained in terms of the Relevance framework, in two experiments we failed to replicate the original causal non-diversity effect. The failure to replicate in Experiment 3 is relatively easy for the Relevance framework to explain: our priming manipulation was so successful that it meant that the information normally made available by manipulations of the premises was lost to participants. However, such an explanation does not work when applied to Experiment 4.

Nonetheless, the absence of a diversity effect in Experiments 3 and 4 does support the Relevance account over similarity-based accounts, even if finding non-diversity effects would have provided stronger evidence. The success of our priming manipulations is also predicted by the Relevance account. Our results are generally consistent with the predictions of the Relevance account but they do call into question some of the effects that originally motivated that account. While some of the effects described in Medin et al.'s original paper have been replicated and extended (see Feeney & Crisp, 2010; Feeney, Shafto, & Dunning, 2007), others, particularly those involving exceptions to diversity, have proven more contentious (see Heit & Feeney, 2005). The various exceptions to diversity documented by Medin et al. (2003) were claimed to be important supporting evidence for the Relevance approach. Because of uncertainty about those effects, it is important that other evidence be found in support of the approach. The effects described here are one such source of evidence. In addition, unlike descriptive models of category-based inductive reasoning, the Relevance approach can be used to motivate predictions about online processing. Thus far, relevance has been manipulated in terms of the likelihood that a particular relation between the categories in an argument will come to mind (see Coley & Vasilyeva, 2010). However, according to Relevance Theory, relevance is determined by the effort that must be expended in order to retrieve or process some new information, *and* by the likely cognitive effects of that information. Thus, as applied to category-based inductive reasoning, the approach allows one to make predictions about processing times and belief revision associated with individual premises. Recent work (Feeney et al., 2010) has demonstrated associations between reading times for individual premises and change to beliefs about argument strength in the light of those premises. Only the Relevance approach allows one to predict these associations.

Implications for Bayesian accounts

The results of our experiments have implications for both of the Bayesian accounts we surveyed in the Introduction. For example, while Heit's Bayesian account of category-based induction inspired both of the methods used in our experiments, and easily predicts the results of Experiments 1 and 2, it has more difficulty accounting for the results of Experiments 3 and 4. The problem for Heit's Bayesian account is not that we have observed effects of causal features, but that the account has no way of specifying what the effect of causal knowledge should be (although such a model could account for any effect on a post hoc basis). Discrepancies between the predictions of a computational-level account, and the performance of the system whose behaviour is to be accounted for, can be expected to arise from time to time and might even tell us something about the right algorithmic-level account. However, when a test arises for which the computational account makes no prediction, the computational-level explanation is clearly in need of extension. Interestingly, the structured Bayesian account by virtue of its context-specific use of different representations over which to calculate the priors (see Kemp & Tenenbaum, 2009), can account for effects of causal knowledge. However, an important question for the structured Bayesian account is what determines the representational structure chosen by any particular participant in a particular setting. Kemp and Tenenbaum (2009, p. 20) claimed that representational structures used for prior determination are selected on the basis of the property in the argument. However, in the experiments described here participants were told only that the property was a biological characteristic. At best, such information might have primed the retrieval of a default tree-structure associated with taxonomic knowledge and it is not clear how the prime (Experiment 3) or some other knowledge (Experiment 4) led some participants to reason about food webs. A related problem is why only some participants reasoned about food webs. For example, more than half of the features listed for causal problems in Experiment 4 were non-causal. The structured Bayesian approach could usefully be modified so that the processes underlying selection of knowledge structures are clearer. One possibility is that the availability of a particular relation (taxonomic, causal or some other relation) between the categories in the argument (see Shafto, Coley, & Vitkin, 2007) determines the representational structure that is chosen for prior estimation.

The structured Bayesian account can also explain the priming effects observed in Experiment 1 because general features are more likely to be transmitted through the taxonomic tree than are idiosyncratic features that are likely to be confined to a branch of the tree. Although the account does not appear to predict the feature listing results a priori, those results can certainly not be presented as evidence against the model, particularly as participants listed features after they had rated argument strength.

Conclusions

The results described here provide evidence for the claim that diversity effects occur, at least partly because diverse premise categories cause participant to bring general and hence highly projectible features to mind. Although our findings are broadly consistent with Relevance and Bayesian approaches, neither approach to category-based induction predicts or perfectly explains all of the results described here. As Bayesian accounts predict, feature distributions play a role in the diversity effect. The Relevance approach to induction also predicts these effects as well as effects of causal primes that are not predicted by Heit's Bayesian approach. However, important questions remain as to how representations are selected in the structured Bayesian approach so that priors may be computed, and about the exceptions to diversity demonstrated by Medin et al. (2003).

Manuscript received 9 September 2010
Revised manuscript received 3 February 2011
First published online 3 May 2011

REFERENCES

Bacon, F. (1620/1898). *Novum Organum*. London: George Bell & Sons.

Blok, S. V., & Gentner, D. (2000). *Reasoning from shared structure. Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Coley, J. D., & Vasilyeva, N. Y. (2010). Generating inductive inferences: Premise relations and property effects. *Psychology of Learning and Motivation*, *53*, 183–226.

Feeney, A. (2007). How many processes underlie category-based induction? Effects of conclusion specificity and cognitive ability. *Memory & Cognition*, *35*, 1830–1839.

Feeney, A., Coley, J., & Crisp, A. K. (2010). The Relevance framework for category-based induction: Evidence from garden path arguments. *Journal of Experimental Psychology: Learning, Memory and Cognition, 36*, 906–919.

Feeney, A., & Crisp, A. K. (2010). A conjunction of fallacies: What different types of causal conjunction error reveal about dual processes for thinking. In K. Manktelow, In D. E. Over, & S. Elqayam (Eds.), *The science of reason: A Festschrift for Jonathan St. B.T. Evans*. Hove, UK: Psychology Press.

Feeney, A., Shafto, P., & Dunning, D. (2007). Who is susceptible to conjunction fallacies in category-based induction? *Psychonomic Bulletin and Review*, *14*, 884–889.

Hahn, U., Bailey, T. M., & Elvin, L. B. C. (2005) Effects of category diversity on learning, memory and generalization. *Memory & Cognition*, *33*, 289–302.

Hayes, B., Goodhew, A., Heit, E., & Gillan, J. (2003). The role of diverse instruction in conceptual change. *Journal of Experimental Child Psychology*, *83*, 256–273.

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 University Press.

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

Heit, E., & Hahn, U. (2001). Diversity based reasoning in children. *Cognitive Psychology, 43*, 243–273.

Heit, E., Hahn, U., & Feeney, A. (2005). Defending diversity. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (pp. 87–99). Washington, DC: American Psychological Association.

Heit, E., & Feeney, A. (2005). Relations between premise similarity and inductive strength. *Psychonomic Bulletin and Review*, *12*, 340–344.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap: A microanalytic review. *Psychonomic Bulletin and Review*, *10*, 785–813.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*, 1, 20–58.

Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory & Cognition*, *31*, 155–165.

Lo, Y., Sides, A., Rozelle, J., & Osherson, D. (2002) Evidential diversity and premise probability in young children's inductive judgement. *Cognitive Science*, *26*, 181–206.

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.

Marr, D. (1982). *Vision*. San Francisco, CA: H. Freeman & Co.

Medin, D., Coley, J. D., Storms, G., & Hayes, B. (2003). A relevance theory of induction. *Psychonomic Bulletin and Review*, *10*, 517–532.

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

Rehder, B. (2007). Property generalization as causal reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches* (pp. 81–113). New York: Cambridge University Press.

Rhodes, M., Brickman, D., & Gelman, S. A. (2008). Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition*, *108*, 543–556.

Rhodes, M., Gelman, S. A., & Brickman, D. (2010). Children's attention to sample composition in learning, teaching and discovery. *Developmental Science*, *13*, 421–429.

Shafto, P., Coley, J. D., & Vitkin, A. (2007). Availability in category-based induction. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental and computational approaches*. Cambridge, UK: Cambridge University Press.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition*, *109*, 175–192.

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

Sloman, S. A. (2005). *Causal models*. Oxford, UK: Oxford University Press.

Sperber, D., & Wilson, D. (1986/1995). *Relevance: Communication and cognition*. Oxford, UK: Basil Blackwell.

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