Memory & Cognition
2007, 35 (7), 1830-1839

How many processes underlie category-based induction? Effects of conclusion specificity and cognitive ability

Aidan Feeney
Durham University, Stockton-on-Tees, England

Two studies investigated participants' sensitivity to the amount and diversity of the evidence when reasoning inductively about categories. Both showed that participants are more sensitive to characteristics of the evidence for arguments with general rather than specific conclusions. Both showed an association between cognitive ability and sensitivity to these evidence characteristics, particularly when the conclusion category was general. These results suggest that a simple associative process may not be sufficient to capture some key phenomena of category-based induction. They also support the claim that the need to generate a superordinate category is a complicating factor in category-based reasoning and that adults' tendency to generate such categories while reasoning has been overestimated.

When we make an inductive inference we use what we already know about the world to make a prediction about some novel event or object. One frequently studied type of inductive inference exploits our tendency to group events and objects into categories. Based on our knowledge about the members of one category, we may be more or less confident that a feature known to be possessed by members of that category will also be possessed by members of another category. Thus, given that we know that dogs possess some property A, we might be relatively confident that wolves also possess the property but very unsure that goldfish have it. There are a variety of factors known to affect the strength of an inductive inference, and a variety of accounts of the operation of these factors have emerged in the literature (for a review see Heit, 2000). In this paper I will describe two individual differences studies of people's sensitivity to some of these factors. The overall aim of these studies was to answer questions about how such sensitivity, where it exists, might best be captured by models of induction.

Category-based induction is most often studied using arguments such as 1 below. The statements above the line are the premises and the statement below the line is the conclusion. Typically, the predicate which participants are asked to project from the categories in the premises to the category in the conclusion is blank. The three characteristics of inductive arguments that I will consider here are the amount and diversity of evidence in the premises and the nature of the conclusion. Consider Arguments 1 and 2. Experimental evidence (e.g., Osherson, Smith, Wilkie, López, & Shafir, 1990; Heit & Feeney, 2005) shows that people consider arguments such as 2 to be stronger than arguments such as 1. This effect is known as the diversity effect and it occurs because people prefer arguments from diverse or dissimilar evidence.

Cows have property *X* (Argument 1)
Horses have property *X*

All mammals have property *X*

Cows have property *X* (Argument 2)
Badgers have property *X*

All mammals have property *X*

It has also been shown that people have more confidence in conclusions that are supported by more evidence (e.g., Osherson et al., 1990; McDonald, Samuels, & Rispoli, 1996). Thus, arguments such as 3 below are perceived to be stronger than those such as 2 or 1. This effect is known as the monotonicity effect and is related to demonstrations in the literature of people's sensitivity to sample size (e.g., Nisbett, Krantz, Jepson, & Kunda, 1983). Finally, the nature of the conclusion may also have an effect on people's category-based inductive inferences. Consider Argument 4 below, which differs from Argument 2 by virtue of the greater specificity of its conclusion. Findings from adults (McDonald et al., 1996) and children (López, Gelman, Gutheil, & Smith, 1992) suggest that the specificity of the conclusion of an argument affects people's inductive inferences. In particular, while López et al. found no sensitivity to diversity and monotonicity in 5-year-old children, they observed sensitivity to these characteristics in 8-year-olds but only for arguments with general conclusions. More recent work (Heit & Hahn, 2001) has succeeded in demonstrating sensitivity to diversity in chil-

A. Feeney, aidan.feeney@durham.ac.uk

dren as young as 5. However, conclusion specificity was not manipulated in that work. At least in children, it appears that particular reasoning phenomena are less likely to be observed for arguments with specific conclusions.

Cows have property *X* (Argument 3)
Horses have property *X*
Badgers have property *X*

All mammals have property *X*

Cows have property *X* (Argument 4)
Badgers have property *X*

All dogs have property *X*

There are several models of category-based induction in the literature (Rips, 1975; Osherson et al., 1990; Sloman, 1993; Smith, Shafir, & Osherson, 1993; McDonald, Samuels, & Rispoli, 1996; Heit 1998; Medin, Coley, Hayes, & Storms, 2003; Shafto, Kemp, Baraff, Coley, & Tenenbaum, 2005). These models differ in terms of their generality, their explanatory level (cf. Marr, 1982) and their degree of precision. The studies to be described here were primarily designed to test two particularly general and well-specified models, both of which were proposed to capture the process(es) involved in induction. According to Osherson et al.'s (1990) similarity coverage model (SCM), in adults there are always at least two processes involved when evaluating a category-based inference. These are the calculation of similarity between the categories in the premises and the category in the conclusion, and calculation of the degree to which the categories in the premises cover the lowest level superordinate category that includes all of the categories in the argument. Similarity is defined as the maximum similarity between the categories in the premises and the conclusion category, while coverage is defined as the average maximum similarity between the categories in the premises and available instances of the inclusive category. According to the model, when the conclusion of an argument is specific, participants must generate an inclusive category. In cases such as these, the SCM posits three processes for category-based induction.

In contrast, Sloman (1993) has described a single process account of inductive reasoning. According to this model, the key phenomena of inductive reasoning may be captured by an associative process designed to calculate the amount of featural overlap between the instances in the premises and conclusion. For example, diversity is explained under this account by diverse premise categories tending to have, on average, greater featural overlap with the category in the conclusion. Similarly, featural overlap is increased when there are more categories in the premises. Thus, where the SCM posits two or three processes, the feature-based model (FBM) suggests that there is only one. In addition, this model does not distinguish, in terms of the number of processes required for argument evaluation, between specific and general conclusions. So, in certain cases, the SCM suggests the involvement of three separate processes whereas the FBM consistently suggests that there is only one.

López et al. (1992) have explained the developmental trends that they observed in children's induction in terms of the SCM. According to them, 5-year-olds do not display sensitivity to diversity and monotonicity because they do not calculate coverage. Instead, they rely on average, rather than maximum, similarity calculations. By age 8, children are held to be able to consider coverage, but they are unable to generate an inclusive category. Hence, although they are sensitive to the amount and diversity of evidence for arguments with general conclusions, they are not similarly sensitive for arguments with specific conclusions. Certainly, the finding that similarity is basic is interesting from an associative point of view. Nonetheless, López et al.'s other findings are problematic for the FBM.

The first aim of the studies to be described here was to examine whether the finding from the developmental literature, that diversity and monotonicity effects are less likely to be observed with specific conclusions (López et al., 1992), will generalize to adults. There is some suggestive evidence that this might be the case. Osherson et al. (1990) observed less sensitivity to both diversity and monotonicity for arguments with specific rather than general conclusions. However, the differences due to conclusion specificity were small, and Osherson et al.'s study contained only one item per condition. In the studies to be described here, participants received six items per condition. If, as might be predicted on the basis of Osherson et al.'s results, adult participants are less likely to display diversity and monotonicity effects with specific conclusions, this would favor the SCM over the FBM.

A second aim of these studies is to examine category-based induction in the context of the argument that there are two processes for thinking (Evans & Over, 1996; Sloman, 1996; Stanovich, 1999). The first type of process (Type 1) is said to be fast, massively parallel, associative and unrelated to working memory while the second type (Type 2) is slow, sequential, symbol manipulating and constrained by general cognitive resources such as working memory. In general, Type 1 processes are said to be sensitive to belief and context, while Type 2 processes are sensitive to abstract structure. Sloman (1996) has argued that category-based induction is a good example of a thinking task achieved primarily (but not always, see Sloman, 1998) via Type 1 processes.

There is converging evidence for a dissociation between types of thinking along these lines (for reviews see Evans, 2003; Osman, 2004). From our current point of view, the most interesting evidence for a dissociation between different processes for thinking comes from work exploiting the difference between the degree to which each process is said to be constrained by general cognitive resources. Stanovich (1999) argues that as Type 2 processes are reliant on general cognitive resources such as working memory whereas Type 1 processes are not, positive associations between measures of cognitive ability and normatively correct performance on primary thinking tasks should be interpreted as suggesting that Type 2 processes are necessary for such performance.

Stanovich's argument has been accepted by many researchers in the literature and there is now a substantial body of work investigating the relationship between cognitive ability and thinking. For example, it is now known that normatively correct performance on a range of tasks ranging from Wason's indicative selection task (Wason, 1966) to Tversky and Kahneman's (1983) conjunction fallacy is predicted by cognitive ability (see Stanovich & West, 1998a, 1998b; Feeney, Shafto, & Dunning, in press). Other phenomena where performance is known to be associated with ability include susceptibility to belief bias in logical reasoning (see Stanovich & West, 1998c; Handley, Capon, Beveridge, Evans, & Dennis, 2004), base rate neglect and pseudodiagnosticity (Stanovich & West, 1998d), and framing effects (Stanovich & West, 1998b). These effects have been extended to reasoning in adolescents where it is known that normative responding is more clearly linked to cognitive ability than is nonnormative responding (Klaczynski, 2001).

In the two studies described here, as well as rating the strength of arguments where evidence characteristics and conclusion specificity had been systematically manipulated, participants also completed a measure of cognitive ability. The discovery of a statistically significant positive relationship between ability and sensitivity to evidence characteristics would suggest that a purely associative account, such as the FBM, is insufficient to explain the key phenomena of induction. It may also be possible, using an individual differences method, to shed light on the veracity and generality of the processes that are posited by the SCM to explain category-based induction. In particular, we might test the model's claim that in order to demonstrate sensitivity to diversity and monotonicity when inductive arguments have specific conclusions, people must generate a superordinate category that contains all of the categories in the argument. If the generation of an inclusive category is an effortful but necessary process for the effect to be reliably observed, then we should find associations between cognitive ability and rates of sensitivity to monotonicity and diversity with specific conclusions.

STUDY 1

Method

Participants. One hundred Durham University students with a mean age of 20 years took part in this study. Of the participants who declared their gender, 19 were male and 80 female.

Design. Participants attempted arguments with premises designed to assess their sensitivity to diversity and monotonicity, with general or specific conclusions. To facilitate an individual differences analysis, I also took a measure of cognitive ability.

Materials. Each participant completed the AH4 (Heim, 1970) a test of cognitive ability that has been used in previous studies of individual differences in thinking (see Newstead, Handley, Harley, Wright, & Farrelly, 2004). There are two 65-item sections to the AH4, each of which is attempted in separate 10-min sessions. The first part is comprised of verbal items concerning direction, verbal opposites, numerical series, verbal analogies, simple arithmetic computations and synonyms. The second part of the test contains diagrammatic items requiring judgments about analogies, sames, subtractions, series and superimpositions. Heim (1970) reported correlations between parts 1 and 2 ranging from .60 to .81. Test-retest consistency and internal reliability for the scale are high (Heim, 1970; Alexopoulos, 1997). Correlations of .60 (Alexopoulos, 1997) and .69 (Heim, 1970) have been found between the AH4 and Raven's matrices.

Each participant completed 48 reasoning problems concerning different types of mammal (all of the reasoning problems used in this study and the next are presented in the appendices). Half of these problems were presented with a specific conclusion (e.g., all foxes, all deer, all tigers) and half with a general conclusion (i.e., all mammals). The problems were constructed in pairs, both members of which shared the same conclusion. This resulted in six items in each of the four conditions of the study. Participants were required to estimate argument strength on a percentage scale.

Half of the problem pairs were designed to measure sensitivity to diversity and the other half were designed to measure sensitivity to amount of evidence. The categories used in the pairs designed to measure sensitivity to diversity were taken from materials described by Osherson et al. (1990). The arguments in each pair of diversity items shared one premise while the second premise was varied. For the diverse item in each pair the category in the second premise was found by Osherson et al. (1990) to be dissimilar to the category in the first premise (the similarity scores reported by Osherson et al. are reported in Appendix A). The second premise in the nondiverse items was found by Osherson and colleagues to be similar to the category in the first premise. For example, Arguments 5 and 6 are, respectively, the diverse and nondiverse members of a pair.

Given the facts that: (Argument 5)
Horses have Property F8.
Seals have Property F8.
How likely is it that:
All foxes have Property F8?
Response (0%–100%): _____%

Given the facts that: (Argument 6)
Horses have Property K2.
Cows have Property K2.
How likely is it that:
All foxes have Property K2?
Response (0%–100%): _____%

Argument pairs designed to test for sensitivity to monotonicity consisted of either two or three premises and a conclusion. Two of the premises in the three-premise arguments were identical to those in the two premise arguments. Arguments 7 and 8 made up a monotonicity pair with a specific conclusion.

Given the facts that: (Argument 7)
Hares have Property A1.
Whales have Property A1.
Grizzly bears have Property A1.
How likely is it that:
All cows have Property A1?
Response (0%–100%): _____%

Given the facts that: (Argument 8)
Hares have Property J1.
Whales have Property J1.
How likely is it that:
All cows have Property J1?
Response (0%–100%): _____%

Because this was an individual differences study it was important that all participants rated the same arguments. Accordingly, we did not rotate content through conditions, and so the premise categories for the each of the conditions were different. Problems were presented in one of four pseudorandom orders constructed so that items from the same pair did not occur consecutively. Finally, all of the arguments concerned blank predicates. Each predicate was a different combination of letters of the alphabet and numbers.

Table 1
Mean Proportions (and Standard Deviations) of Pairs (Out of Six) From Experiment 1 Displaying Each Effect As a Function of Reasoning Phenomenon and Conclusion Type

Conclusion Type	Diversity Items		Monotonicity Items	
	M	*SD*	*M*	*SD*
General	.58	.27	.66	.24
Specific	.46	.22	.50	.25
Average	.52		.58	

Procedure. Participants completed the reasoning problems and ability measure in large groups during lectures.

Results

Overall sensitivity to evidence characteristics. The results were coded by problem pair. The percentage of participants displaying each effect for each problem pair is reported in Appendix A. For the diversity items, I analyzed the proportion of problem pairs on which participants gave higher ratings of argument strength to arguments with diverse premises. For the monotonicity pairs I examined the proportion of pairs where participants gave higher ratings of argument strength to arguments with three premise categories.

Two dependent measures *t* tests revealed significant effects of conclusion type for diversity items [$t(99) = 4.05$, $p < .001$] and for monotonicity items [$t(99) = 4.99$, $p < .001$]. The means from these analyses are presented in Table 1 where it may be seen that overall, we have obtained stronger evidence of sensitivity to both characteristics of the evidence for problems with general conclusions than for those with specific conclusions. When the conclusion category was general, single sample *t* tests revealed above chance responding for monotonicity items [$t(99) = 6.79$, $p < .001$] and diversity items [$t(99) = 2.83$, $p < .01$]. However, when the conclusion category was specific, responding did not differ from chance for monotonicity items [$t(99) = .133$, $p > .5$]. Sensitivity to diversity with specific conclusions was close to being significantly less than chance [$t(99) = -1.95$, $p < .06$].

Individual differences analysis. The second analysis examined the relationship between scores on the cognitive ability measure and people's sensitivity to diversity and sample size. Overall performance on the AH4 ($M = 95.6$, $SD = 14.40$) was close to the norm for a university sample ($M = 96.4$, $SD = 15.01$). Collapsed across conclusion, there were significant correlations between ability and sensitivity to diversity ($r = .27$, $p < .007$) and between ability and sensitivity to monotonicity ($r = .24$, $p < .02$). When collapsed across phenomenon, while there was a significant association between sensitivity to characteristics of the evidence and cognitive ability for problems with general conclusions ($r = .33$, $p < .001$), the association fell just short of statistical significance for problems with specific conclusions ($r = .19$, $p < .07$).

Detailed analysis of the relationship between performance on the AH4 results for each of the experimental conditions confirmed the pattern of results described above. Thus, for arguments with general conclusions there were significant associations between AH4 scores and sensitivity to diversity ($r = 27$, $p < .008$) and sensitivity to monotonicity ($r = .24$, $p < .02$). However, the corresponding associations for arguments with specific conclusions were nonsignificant ($r = .16$, $p > .11$ and $r = .12$, $p > .23$, respectively).

Discussion

The results of Study 1 contain several noteworthy findings. In particular, although rates of sensitivity to diversity and monotonicity were above chance, this was only true for arguments with general rather than specific conclusions. Adults, in much the same fashion as young children (López et al., 1992), appear to have problems with specific conclusions. Given that Osherson et al.'s (1990) original demonstration of a monotonicity effect for specific conclusions involved only one argument, and that López et al. did not include an adult control condition in their experiments, perhaps adult sensitivity to monotonicity and diversity in arguments with specific conclusions has been overstated. I will return to this issue in the next study.

A second finding of note is the positive and significant association between cognitive ability and sensitivity to diversity and monotonicity for arguments with general conclusions. As measures of cognitive ability such as the AH4 are held to tax memory resources, and as associative Type 1 processes are supposed not to draw on such resources, one interpretation of this finding is that a simple associative account is insufficient to explain people's sensitivity to certain inductive reasoning phenomena. However, although significant, the correlation coefficients reported here are small, and care should be exercised in their interpretation. I will return to these interpretational issues in the general discussion.

STUDY 2

In Study 2, I attempted to replicate the results of Study 1. In the first study, I used different sets of categories in each of the four conditions testing for diversity and monotonicity phenomena. In Study 2, I attempted to achieve more control over the materials by creating sets of items which could be used to manipulate evidential diversity, number of premises, and conclusion specificity.

Method

Participants. Twenty-two male and 93 female students of psychology at Durham University participated in this study. Their mean age was 19 years.

Design. The study had the same repeated measure design as was used in Study 1. To facilitate an individual differences analysis, participants also completed the AH4.

Materials. Participants completed 36 reasoning problems instead of 48 (see Appendix B for full list of problems used in the experiment). Six problem sets were generated, each containing six individual problems. Each set employed three premise categories (e.g., mice, dolphins, squirrels). The item in each set that was designed to test for sensitivity to number of premises involved presentation of all three categories in the premise of the argument. Diversity items involved presentation of just two, either a pair found by Osherson et al. to be similar, or a pair found to be dissimilar. Specific argu-

Table 2
Mean Proportions (and Standard Deviations) of Pairs (Out of Six) From Experiment 2 Displaying Each Effect As a Function of Reasoning Phenomenon and Conclusion Type

			Monotonicity Items			
	Diversity Items		Compared to Diverse Argument		Compared to Nondiverse Argument	
Conclusion Type	*M*	*SD*	*M*	*SD*	*M*	*SD*
General	.59	.27	.68	.23	.82	.21
Specific	.52	.25	.59	.23	.69	.23
Average	.56		.64		.76	

ments were formed by using a specific category in the conclusion (e.g., all dogs). The same specific conclusion category was used for all of the specific problems formed from any one set. Mammal was the conclusion category for all general arguments. As was the case in Study 1, participants were asked to provide a percentage strength rating for each argument.

Procedure. Participants completed the AH4 and the reasoning task during the course of a lecture.

Results

Overall sensitivity to evidence characteristics. Each set of three arguments was presented with a general or specific conclusion. For each conclusion type, I coded the proportion of sets where the diverse argument was rated stronger than the less diverse argument. When coding for sensitivity to monotonicity, I coded the proportion of sets where the three premise argument was rated stronger than the diverse two premise argument, and the proportion of sets where the three premise argument was rated stronger than the nondiverse argument.

The mean proportion of sets displaying each effect, broken down by Conclusion, is displayed in Table 2. Dependent measures *t* tests revealed an effect of Conclusion for diversity items [$t(114) = 2.96$, $p < .005$], for monotonicity items compared to diverse arguments [$t(114) = 3.06$, $p < .003$], and for monotonicity items compared to nondiverse arguments [$t(114) = 6.03$, $p < .001$].

Single sample *t* tests revealed above chance sensitivity to diversity when the conclusion was general [$t(114) = 3.59$, $p < .001$], but not when it was specific [$t(114) = .82$, $p = .41$]. Sensitivity to monotonicity for arguments with general conclusions was significantly above chance whether sensitivity to monotonicity was defined as higher ratings for three premise arguments than for diverse arguments [$t(114) = 8.48$, $p < .001$], or as higher ratings for the three premise arguments than for nondiverse arguments [$t(114) = 16.50$, $p < .001$]. For specific arguments, sensitivity to monotonicity was also significantly above chance, again whether monotonicity was measured with reference to diverse arguments [$t(114) = 3.99$, $p < .001$], or to nondiverse arguments [$t(114) = 8.65$, $p < .001$].

Individual differences analysis. I analyzed the relationship between people's performance on the AH4 and their sensitivity to diversity and monotonicity. The mean AH4 score in this study was 106.3 ($SD = 10.65$). This is significantly higher ($p < .001$) than the published norm (see Heims, 1970) of 96.4 ($SD = 15.01$, $N = 726$) for a university sample. Thus, we appear to have sampled a higher part of the distribution of AH4 scores in this study than we did in Study 1. This may be because the samples from Studies 1 and 2 are made up of students taking different degrees at Durham University. As there are considerable differences in the academic entry requirements for some degrees, we might expect to observe corresponding differences in performance on measures of cognitive ability. Because there was substantially less variability about the mean AH4 score in this study than is the norm for the test, I report correlation coefficients that have been adjusted for restricted variance.[1]

Collapsed across conclusion type we observed a significant correlation between the tendency to be sensitive to diversity and AH4 scores [$r(\text{adj}) = .29$, $p < .005$]. However, the overall association between sensitivity to monotonicity and cognitive ability only reached statistical significance [$r(\text{adj}) = .24$, $p < .01$], when sensitivity to monotonicity was assessed relative to nondiverse arguments. When assessed relative to diverse arguments, sensitivity to monotonicity was not significantly associated with ability [$r(\text{adj}) = .08$].

When we examine the relationship between cognitive ability and performance in each of our experimental conditions, we find significant associations for diversity items with general conclusions [$r(\text{adj}) = .23$, $p < .02$], and for diversity items with specific conclusions [$r(\text{adj}) = .25$, $p < .01$]. For arguments with general conclusions, sensitivity to monotonicity was associated with cognitive ability whether such sensitivity was coded relative to diverse arguments [$r(\text{adj}) = .22$, $p < .02$], or relative to nondiverse arguments [$r(\text{adj}) = .26$, $p < .005$]. For specific arguments, associations with cognitive ability were weaker both when sensitivity to monotonicity was coded relative to diverse arguments [$r(\text{adj}) = .1$], or relative to nondiverse arguments [$r(\text{adj}) = .14$].

To explore the significant correlation between cognitive ability and sensitivity to diversity on arguments with specific conclusions, we present the proportion of trials on which sensitivity was observed among participants in each decile of the distribution of AH4 scores from Stud-

Table 3
Mean AH4 Scores and Mean Proportions of Specific Conclusion Pairs (Out of Six) for Each Decile of AH4 Scores From Studies 1 and 2, Where the Diverse Argument Was Rated Stronger Than the Nondiverse Argument

	Study 1			Study 2		
Decile	*N*	AH4	DS	*N*	AH4	DS
1st	10	67	.42	11	83	.44
2nd	10	79	.43	12	96	.43
3rd	10	87	.37	11	102	.53
4th	10	91	.48	12	105	.61
5th	10	96	.47	11	107	.53
6th	10	101	.45	12	109	.47
7th	10	103	.42	11	110	.48
8th	10	106	.42	12	114	.53
9th	10	111	.55	11	117	.52
10th	10	116	.57	12	120	.64

Note—AH4, mean AH4 score; DS, diversity sensitivity.

ies 1 and 2. As might be expected given the size of the relevant correlation coefficients, the data are somewhat noisy. Nonetheless, an examination of Table 3 suggests some patterns. First, the decile scores suggest that in both studies, participants lowest in ability were sensitive to diversity at below chance levels. Because the higher ability participants in the 9th and 10th deciles of Study 1 were considerably more sensitive to diversity than were the lower ability participants in the 1st, 2nd, and 3rd deciles, the data for Study 1 is consistent with the predicted association between sensitivity to diversity with specific conclusions and cognitive ability. However, an important difference between the studies is due to the highest ability group in Study 2 who are sensitive to diversity on almost two thirds of arguments with specific conclusions. This group stretches the difference in sensitivity scores between the most and least able participants. That such high rates of sensitivity are not observed in Study 1, suggests that it may be this very high ability group that is responsible for the significant correlation observed in Study 2 between cognitive ability and sensitivity to diversity on specific arguments.

Discussion

The results of Study 2 confirm and extend the results of Study 1. Once again we have observed above chance rates of sensitivity to evidential diversity for arguments with general conclusions. There is also stronger evidence in this study than in the previous study of sensitivity to monotonicity for both types of conclusions, regardless of whether one codes for such sensitivity by comparing the three premise argument to the diverse or to the nondiverse argument. In addition, for arguments with general conclusions, sensitivity to the diversity and amount of evidence were both associated with cognitive ability.

A novel finding in this study is that for specific arguments, sensitivity to diversity (but not to amount of evidence) is positively associated with cognitive ability. It is likely that we have observed such an association in this study but not in Study 1 because of the difference between the cognitive ability of the participants in each study. Recall that although the AH4 scores of participants in Study 1 were substantially lower than the scores of participants in Study 2, comparison across both studies of the AH4 scores by decile, reveals that sensitivity to diversity in the lowest ability groups was approximately equal. However, there was a very high ability subgroup in Study 2 whose members were sensitive to diversity on more arguments with specific conclusions than were participants in any of the other 19 decile groups across the two studies. Accordingly, the difference between the highest and lowest ability groups was larger in Study 2 than it was in Study 1, which may help to explain why the correlation with ability was significant in the former case but not in the latter.

Newstead et al. (2004) report a similar finding in a study of Wason's selection task. Stanovich and West (1998a) reported that SAT scores were associated with the tendency to give the logically correct response on indicative versions of Wason's task (concerning rules about what is or was the case) but not with logical responding on deontic versions of the task (what ought to or should be the case). Newstead et al. observed this pattern with very high ability participants only. In studies with participants of moderate ability, the reverse pattern was observed. These authors also explain their findings in terms of the ability levels of their participants: phenomena that are very cognitively demanding will only be observed with a sample containing sufficient numbers of high ability participants. I will return to the implications of this finding for the SCM in the General Discussion.

Finally, when assembling the materials for Studies 1 and 2, we used similarity ratings from Osherson et al. (1990) to select the premise categories for diverse and nondiverse arguments. One possible objection to the results of those studies is that for the diversity arguments with specific conclusions, we did not pretest the similarity between each of the premise categories and the specific conclusion category (this objection does not apply to the monotonicity arguments presented with specific conclusions, as the SCM always predicts monotonicity with the addition of an extra premise category). According to the SCM, the strength of a two premise argument with a specific conclusion is comprised of a coverage score and a similarity score. Because we used Osherson et al.'s similarity data to select premise pairs, we have good reason for supposing that the coverage score will be higher for the more diverse item of each pair. However, we did not control for differences between similarity scores. The similarity score is given by the maximum similarity between the premise categories and the conclusion category. Consider Arguments 5 and 6. If horses are more similar to foxes than are either seals or cows, then for each argument the similarity score is equal to the similarity between horses and foxes. However, it is possible that seals or cows are more similar to foxes than are horses, and in this case, Arguments 5 and 6 would have different similarity scores. Perhaps, the lack of an overall diversity effect in Studies 1 and 2 for specific arguments, might be due to stronger similarity scores for the nondiverse pair than for the diverse pair.

To rule out the possibility of a confound, 67 undergraduate participants at Durham University rated, on a scale from 1–9, the similarity between 20 pairs of categories. Each pair was comprised of a specific conclusion from a diverse/specific argument used in Studies 1 and 2, and one of the premise categories from that argument. The pairs were presented in one of two randomly determined orders and the categories in each pair appeared in the order that they had occurred in the argument. The results of this posttest are presented in Table 4 where it may be seen that for six of the seven items used in the two studies, the category common to both premise sets is rated more similar to the conclusion category than either of the other premise categories in the item. Because the SCM uses MAX similarity to derive the similarity measure, it predicts that the similarity measure for both of the arguments in these six items should be identical.

The single item where the common category was rated less similar to the conclusion category than were either of

Table 4
Mean Similarity Ratings, By Item, for the Categories Used in the Specific Arguments of Studies 1 and 2 to Test for Sensitivity to Diversity

Premise Category	Conclusion Category	Similarity of Category to Conclusion		
		Common	Diverse	Nondiverse
Horse seals Horse cows	foxes	3.69	1.94	2.77
Cows squirrels Cows elephants	deer	1.86	2.67	2.11
Cows dolphins Cows rhinos	dogs	3.13	2.56	1.64
Chimps elephants Chimps gorillas	goats	4.73	2.64	2.69
Gorillas mice Gorillas rhinos	wolves	2.91	2.42	2.80
Squirrels seals Squirrels chimps	tigers	3.09	2.92	1.94
Mice dolphins Mice squirrels	dogs	3.58	2.42	2.63

Note—In Study 2, the third item in the table was replaced by the final item. The leftmost number in each row is the similarity score reported by Osherson et al. (1990) for the premise pair in that row.

the other two categories, was the *cow, squirrels/elephants therefore deer* item. However, as squirrels (the second premise category in the diverse argument) were perceived to be more similar to deer than were elephants (the second category in the nondiverse argument), this should have had the effect of making a diversity effect more, rather than less, likely. However, this item was least likely to produce a diversity effect in Study 1. In sum, the results of the posttest suggest that the low rates of sensitivity to diversity observed in Studies 1 and 2 cannot be attributed to systematic biases caused by greater maximum similarity between the conclusion and premise categories in the nondiverse argument.

GENERAL DISCUSSION

The results that have been presented here help us to evaluate two claims that have been made in the literature on induction. First, they are problematic for the claim that there is no psychological difference between the evaluation of specific and general category-based inductive arguments (see Sloman, 1993). In both studies we have observed less sensitivity to the characteristics of the evidence when the conclusion was specific rather than general. This finding confirms a prediction derived from the SCM and is consistent with developmental findings showing that sensitivity to diversity and monotonicity emerges for general arguments before it emerges for specific arguments (see López et al., 1992).

However, another way to characterize our findings is that we have observed almost no sensitivity to diversity for arguments with specific conclusions. Viewed in this light, our results are problematic for the SCM, as it appears to predict a phenomenon that is not observed. However, in Study 2 we did observe an association between ability and sensitivity to diversity for specific arguments, although no association was observed in Study 1, where the ability of the sample was lower. The correlation observed in Study 2 is important because it suggests that some very able participants were consistently displaying sensitivity to the phenomenon. So it is not the case the sensitivity to diversity is never observed for specific arguments, although it does appear to be considerably rarer in adults than was initially supposed. However, it is possible that participants had difficulty with the diversity/specific items because they rated the arguments in random order. This may have increased the difficulty of the task, relative to the forced choice method often used.

The second claim to which our results are relevant is that category-based induction is primarily achieved via the operation of an associationist system (Sloman, 1993, 1996). There is a strong argument in the literature on dual processes in reasoning that associations with cognitive ability suggest the involvement of Type 2 processes. Such processes are said to be explicit, slow, sequential, symbol manipulating and thus, limited by basic cognitive constraints such as working memory. Type 1 processes, by contrast, are held to be fast, massively parallel, associative, and as a result, relatively independent of constraints such as memory. As working memory is known to be very important in determining performance on tests of cognitive ability (e.g., Carpenter, Just, & Shell, 1990), our finding of an association between a measure of ability and sensitivity to certain inductive phenomena suggests that the tendency to display those phenomena is, to some extent at least, dependent on Type 2 processes. It is important to state here that this finding does not question the view that reasoning based on similarity is often dependent on associative Type 1 processes. There are several sources of evidence suggesting that such reasoning is basic. For example, it appears to occur across cultures (see López, Atran, Coley, Medin, & Smith, 1997) and in very young children (López et al., 1992). However, our results suggest that perhaps the diversity and monotonicity effects rely on more than simple similarity calculations.

There are a number of reasons for exercising caution when interpreting the individual differences findings from this experiment. First, the data is correlational, and the correlations we have reported, although statistically significant, are not large. It is possible that the statistical variance shared by the ability measure and the reasoning task, is attributable to some factor other than general cognitive ability. That is, these correlations may not be indicative of the involvement of Type 2 reasoning processes in induction. However, in other areas of the literature, relatively weak, but statistically significant correlational data has subsequently been supported by the results of experimental studies employing a variety of experimental manipulations designed to differentially affect Type 1 and Type 2 processes (see De Neys, 2006; Evans & Curtis-Holmes, 2005). Before strong claims can be made about the involvement of Type 2 reasoning processes in category-based induction, experimental studies of the effects on reasoning of secondary tasks and time pressure are essential.

Another difficulty is that for specific arguments in Study 2 we observed a significant association between

cognitive ability and sensitivity to diversity but not sensitivity to monotonicity. A complication in interpreting these results is that sensitivity to monotonicity in Study 2 could be measured relative to diverse or nondiverse arguments. Overall, participants showed greater sensitivity to monotonicity when it was measured relative to nondiverse arguments. In addition, cognitive ability was associated with overall sensitivity to monotonicity only when it was measured relative to nondiverse arguments. Both of these findings suggest that at least some participants may have been consistently rating diverse arguments and three premise arguments stronger than nondiverse arguments, without regard to the relative strength of three premise and diverse arguments. Because they were more likely to be sensitive to diversity, this pattern of responding is likely to have been more prevalent among high ability participants. This may explain the absence of a significant association between cognitive ability and overall sensitivity to monotonicity, when sensitivity to monotonicity was coded relative to diverse arguments. The pattern of responding I have described may also have also have interfered with the detection of a significant association between cognitive ability and sensitivity to monotonicity with specific arguments.

Leaving aside the dissociation between sensitivity to diversity and monotonicity observed for specific arguments in Study 2, the individual differences results are consistent with a number of more recent accounts of category-based induction which have taken a hypothesis-testing approach to argument evaluation (see Heit, 1998; McDonald et al., 1996; Medin et al., 2003). For example, Heit's Bayesian account predicts that people should be sensitive to evidence characteristics in a normatively appropriate manner. Where there are individual differences in such sensitivity it appears to predict that the most able individuals in any sample should be most likely to display sensitivity. The relevance approach to induction (Medin et al., 2003) might explain correlations between ability and sensitivity to evidence characteristics in terms of more able participants being more likely to select the relevant relation between the categories in the argument in order to guide their hypotheses about the blank predicate.

In conclusion, the studies described here demonstrate that sensitivity to diversity and monotonicity are predicted by cognitive ability but that such sensitivity is comparatively rare in the presence of specific conclusions. These findings suggest that an associative account may be insufficient to capture the full range of inductive phenomena. They also suggest that people's sensitivity to evidence characteristics in the presence of specific conclusions may have been overestimated. However, had a measure of cognitive ability not been included in these studies, we would have been forced to conclude that they revealed no sensitivity at all to evidential diversity for arguments with specific conclusions. In future it is to be hoped that researchers will ask participants to provide data that will allow for such analyses, as asking who does what in reasoning experiments can be a very fruitful question.

AUTHOR NOTE

Correspondence regarding this article should be addressed to A. Feeney, Department of Psychology, Durham University Queen's Campus, Thornaby, Stockton-on-Tees TS17 6BH, England (e-mail: aidan.feeney@durham.ac.uk).

REFERENCES

ALEXOPOULOS, D. S. (1997). Reliability and validity of Heim's AH4 in Greece. *Personality & Individual Differences*, **22**, 429-432.

CARPENTER, P. A., JUST, M. A., & SHELL, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven Progressive Matrices test. *Psychological Review*, **97**, 404-431.

DE NEYS, W. (2006). Automatic-heuristic and executive-analytic processing in reasoning: Chronometric and dual task considerations. *Quarterly Journal of Experimental Psychology*, **59**, 1070-1100.

EVANS, J. ST. B. T. (2003). In two minds: Dual process accounts of reasoning. *Trends in Cognitive Sciences*, **7**, 454-459.

EVANS, J. ST. B. T., & CURTIS-HOLMES, J. (2005). Rapid responding increases belief bias: Evidence for the dual process theory of reasoning. *Thinking & Reasoning*, **11**, 382-389.

EVANS, J. ST. B. T., & OVER, D. E. (1996). *Rationality and reasoning*. Hove, U.K.: Psychology Press.

FEENEY, A., SHAFTO, P., & DUNNING D. (in press). Who is susceptible to conjunction fallacies in category-based induction? *Psychonomic Bulletin & Review*.

HANDLEY, S. J., CAPON, A., BEVERIDGE, M., DENNIS, I., & EVANS, J. ST. B. T. (2004). Working memory and inhibitory control in the development of children's reasoning. *Thinking & Reasoning*, **10**, 175-196.

HEIM, A. W. (1970). *AH4 group test of intelligence* [Manual]. London: National Foundation for Educational Research.

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

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

HEIT, E., & FEENEY, A. (2005). Relations between premise similarity and inductive strength. *Psychonomic Bulletin & Review*, **12**, 340-344.

HEIT, E., & HAHN, U. (2001). Diversity based reasoning in children. *Cognitive Psychology*, **43**, 243-273.

KLACZYNSKI, P. A. (2001). Analytical and heuristic processing influences on adolescent reasoning and decision making. *Child Development*, **72**, 844-861.

LÓPEZ, A., ATRAN, S., COLEY, J. D., MEDIN, D. L., & SMITH, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, **32**, 251-295.

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

MARR, D. (1982). *Vision*. San Francisco, CA: Freeman.

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

MEDIN, D., COLEY, J. D., STORMS, G., & HAYES, B. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, **10**, 517-532.

NEWSTEAD, S. E., HANDLEY, S. J., HARLEY, C., WRIGHT, H., & FARRELLY, D. (2004). Individual differences in deductive reasoning. *Quarterly Journal of Experimental Psychology*, **57**, 33-60.

NISBETT, R. E., KRANTZ, D. H., JEPSON, D., & KUNDA, Z. (1983). The use of statistical heuristics in everyday reasoning. *Psychological Review*, **90**, 339-363.

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

OSMAN, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, **11**, 988-1010.

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

SHAFTO, P., KEMP, C., BARAFF, E., COLEY, J. D., & TENENBAUM, J. B.

(2005). Context sensitive induction. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 2003-2008) Mahwah, NJ: Erlbaum.

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

SLOMAN, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, **119**, 3-22.

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

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

STANOVICH, K. E. (1999). *Who is rational: Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

STANOVICH, K. E., & WEST, R. F. (1998a). Cognitive ability and variation in selection task performance. *Thinking & Reasoning*, **4**, 193-230.

STANOVICH, K. E., & WEST, R. F. (1998b). Individual differences in framing and conjunction effects. *Thinking & Reasoning*, **4**, 289-317.

STANOVICH, K. E., & WEST, R. F. (1998c). Individual differences in rational thought. *Journal of Experimental Psychology: General*, **127**, 161-188.

STANOVICH, K. E., & WEST, R. F. (1998d). Who uses base rates and P(D/~H)? An analysis of individual differences. *Memory & Cognition*, **26**, 161-179.

TABACHNICK, B. G., & FIDELL, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

TVERSKY, A., & KAHNEMAN, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, **90**, 293-315.

WASON, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (Vol. 1, pp. 131-151). Harmondsworth, U.K.: Penguin.

NOTE

1. Correlation coefficients were adjusted using the following formula (see Tabachnick & Fidell, 1996):

$$\tilde{r}_{XY} = \frac{r_{t(XY)}\left[\frac{S_X}{S_{t(X)}}\right]}{\sqrt{1 + r^2_{t(XY)}\left[\frac{S_X}{S_{t(X)}}\right] - r^2_{t(XY)}}},$$

where $\tilde{r}_{XY}$ = the adjusted correlation, $r_{t(XY)}$ = the correlation, S_X = standard deviation of the truncated variable X, and $S_{t(X)}$ = standard deviation of (observed) truncated variable X.

APPENDIX A
Materials Used in Study 1

Diversity Items

Diverse–Specific	Nondiverse–Specific	Conclusion	Percentage
horses seals (.37)	horses cows (.93)	foxes	55
cows squirrels (.49)	cows elephants (.79)	deer	37
cows dolphins (.26)	cows rhinos (.79)	dogs	46
chimps elephants (.53)	chimps gorillas (.97)	goats	55
gorillas mice (.37)	gorillas rhinos (.65)	wolves	43
squirrels seals (.27)	squirrels chimps (.65)	tigers	38
Diverse–General*	Nondiverse–General		
horses dolphins (.33)	horses elephants (.80)		50
gorillas seals (.34)	gorillas elephants (.65)		47
mice dolphins (.17)	mice squirrels (.94)		72
dolphins elephants (.29)	dolphins seals (.92)		71
rhinos seals (.32)	rhinos elephants (.92)		55
cows seals (.38)	cows gorillas (.59)		51

Monotonicity Items

Two Cases–Specific	Three Cases–Specific		
horses wolves	horses wolves hippos	dogs	32
zebras koalas	zebras koalas hyenas	hippos	51
hares whales	hares whales grizzly bears	cows	58
dogs cows	dogs cows koalas	boars	52
boars foxes	boars foxes tigers	horses	54
hippos deer	hippos deer foxes	grizzly bears	55
Two Cases–General	Three Cases–General		
boars koalas	boars koalas horses		69
goats dogs	goats dogs zebras		61
wolves cows	wolves cows koalas		65
tigers grizzly bears	tigers grizzly bears horses		76
hares hyenas	hares hyenas boars		63
whales deer	whales deer tigers		64

Note—The numbers in parentheses are the similarity scores reported by Osherson et al. (1990) for each premise pair. Percentage refers to the percentage of participants displaying the effect of interest on each item. *In all cases, the general conclusion category was mammals.

APPENDIX B
Materials Used in Study 2

Premises			Specific	% Participants Displaying Effect of Interest on Each Item					
				Monotonicity					
				General*		Specific		Diversity	
Diverse	Nondiverse	Three-Category	Conclusion	D	ND	D	ND	General	Specific
horses seals (.37)	horses cows (.93)	horses seals cows	foxes	63	87	58	60	71	48
cows squirrels (.49)	cows elephants (.79)	cows squirrels elephants	deer	73	80	50	63	53	63
mice dolphins (.17)	mice squirrels (.94)	mice dolphins squirrels	dogs	60	83	62	76	70	55
chimps elephants (.53)	chimps gorillas (.97)	chimps elephants gorillas	goats	63	87	49	81	73	71
gorillas mice (.37)	gorillas rhinos (.65)	gorillas mice rhinos	wolves	75	83	70	70	44	42
squirrels seals (.27)	squirrels chimps (.56)	squirrels seals chimps	tigers	74	72	63	63	44	33

Note—The numbers in parentheses are the similarity scores reported by Osherson et al. (1990) for each premise pair. D refers to the percentage of participants displaying the effect when strength ratings for the three-premise argument were compared to ratings for the diverse two-premise argument, and ND refers to the percentage of participants displaying the effect when the three-premise argument was compared to the nondiverse two-premise argument. *In all cases, the general conclusion category was mammals.

(Manuscript received August 3, 2006;
revision accepted for publication January 26, 2007.)