

Relations between premise similarity and inductive strength

EVAN HEIT

University of Warwick, Coventry, England

and

AIDAN FEENEY

University of Durham, Durham, England

According to the diversity principle, diverse evidence is strong evidence. There has been considerable evidence that people respect this principle in inductive reasoning. However, exceptions may be particularly informative. Medin, Coley, Storms, and Hayes (2003) introduced a relevance theory of inductive reasoning and used this theory to predict exceptions, including the nondiversity-by-property-reinforcement effect. A new experiment in which this phenomenon was investigated is reported here. Subjects made inductive strength judgments and similarity judgments for stimuli from Medin et al. (2003). The inductive strength judgments showed the same pattern as that in Medin et al. (2003); however, the similarity judgments suggested that the pattern should be interpreted as a diversity effect, rather than as a nondiversity effect. It is concluded that the evidence regarding the predicted nondiversity-by-property-reinforcement effect does not give distinctive support for relevance theory, although this theory does address other results.

Why do some observations lead to broad generalizations, whereas other observations do not have as much influence on people's beliefs? One principle for evaluating evidence is the diversity principle, which states that more diverse evidence should lead to stronger inferences than a narrow sample of evidence does. This principle has been influential in the history of science, capturing scientists' preference for testing a theory with a diverse set of experiments, rather than repeatedly conducting the same experiment or close replications (e.g., Salmon, 1984). There has also been a widespread effort by psychologists to document how diversity of evidence affects the way people carry out a variety of cognitive activities. Thus far, sensitivity to diversity has been found in hypothesis testing (Kincannon & Spellman, 2003; López, 1995) and diagnostic reasoning (Kim & Keil, 2003), as well as in children's reasoning about the physical world (Hayes, Goodhew, Heit, & Gillan, 2003). Furthermore, there is well-established evidence from categorization research that more variable observations promote broader or stronger generalizations (e.g., Fried & Holyoak, 1984; Homa & Vosburgh, 1976; Posner & Keele, 1968).

In an influential study, Osherson, Smith, Wilkie, López, and Shafir (1990) documented diversity effects in adults'

informal inductive reasoning by using written arguments such as the following:

- (1) Hippos require Vitamin K for the liver to function.
Rhinos require Vitamin K for the liver to function.

All mammals require Vitamin K for the liver to function.
- (2) Hippos require Vitamin K for the liver to function.
Hamsters require Vitamin K for the liver to function.

All mammals require Vitamin K for the liver to function.

People judged arguments such as (2) to be stronger than arguments such as (1), showing sensitivity to the greater diversity of the premise categories in Argument 2 (hippos and hamsters).

In general, there has been a great deal of evidence of adults, particularly Western college students, following the diversity principle in inductive reasoning (see Heit, 2000, for a review; and see Heit & Hahn, 2001, for a consideration of developmental work). The main exceptions to the diversity effect seem to be due the use of other knowledge, such as expertise, rather than diversity (e.g., López, Atran, Coley, Medin, & Smith, 1997; Proffitt, Coley, & Medin, 2000). So these exceptions do not invalidate the diversity principle so much as show that it can be overridden by other knowledge (Heit, Hahn, & Feeny, 2005). Indeed, the diversity effect is considered one of the touchstone results explained by previous models of inductive reasoning (Heit, 1998; Osherson et al., 1990; Sloman, 1993).

In a recent paper, Medin, Coley, Storms, and Hayes (2003) reported further exceptions to the diversity principle. Some of these exceptions involve diversity's being

A preliminary report of this study appeared in Heit, Hahn, and Feeny (2005). The authors are grateful to Ian Davison and Yana Weinstein for assistance with this research. Please address correspondence to E. Heit, Department of Psychology, University of Warwick, Coventry CV4 7AL, England (e-mail: e.heit@warwick.ac.uk).

overridden by other knowledge, particularly causal knowledge. However, one phenomenon makes a more direct challenge. In the nondiversity-by-property-reinforcement effect, “if an otherwise diverse set of premises shares a salient property not shared by the conclusion category, the reinforcement of the property might weaken that argument relative to a related argument with less diverse premises” (p. 523). This phenomenon is illustrated by the following example:

(3) Penguins have property X.
Eagles have property X.

All animals have property X.

(4) Penguins have property Y.
Polar bears have property Y.

All animals have property Y.

When given a forced choice, subjects judged the two animals that are taxonomically near (penguins and eagles) to be more similar than the two animals that are taxonomically distant (penguins and polar bears). However, when subjects were asked to assess the inductive strength of each argument, Argument 4 was judged to be less convincing than Argument 3. That is, the premise categories in Argument 3 (penguins and eagles) appeared to describe less diverse evidence; yet this was the stronger argument. Intuitively, although penguins and polar bears are taxonomically distant, they still share the property of living in a cold climate. It might seem that Property Y does not extend to all animals but applies only to animals in cold climates.

The theoretical importance of the nondiversity-by-property-reinforcement effect is that it could not be explained by previous models of inductive reasoning. Instead, such a result would provide distinctive evidence for Medin et al.’s (2003) own account, relevance theory, which gives an explanation in terms of people looking for distinctive properties of premise categories, on the assumption that these categories were not chosen randomly but, instead, were presented as part of a discourse (see also Tenenbaum & Griffiths, 2001). Relevance theory refers to listeners drawing inferences about speakers’ communicative goals against the context of the discourse and shared background knowledge.

Medin et al. (2003) investigated this effect, using several stimulus sets, and overall did find significant evi-

dence. However, the results were not always consistent. Sometimes, the similarity comparisons went in the direction opposite to that of the anticipated one—that is, sometimes taxonomically near animals were judged to be more diverse than taxonomically distant animals. On other occasions, the inductive strength judgments went in the direction of diversity, rather than nondiversity. Hence, we conducted a further test of this phenomenon, with the aim of determining whether relevance theory should be favored over previous accounts of inductive reasoning. We used Medin et al.’s (2003) stimuli and, in general, followed their procedure, except for collecting similarity judgments in a different way. Rather than asking subjects to make a forced choice between two taxonomically near animals and two taxonomically distant animals, we asked subjects to make individual similarity ratings corresponding to each of the arguments. This procedure facilitated the key analysis, in which the correlation between similarity and inductive strength was examined, allowing consideration of the whole pattern of results.

METHOD

Subjects

Two groups of volunteers were recruited on the University of Durham’s Queen’s Campus. There were 72 subjects who made judgments of inductive strength and 45 subjects who made similarity ratings. Within the induction condition, there were two subgroups: Half were asked to give justifications for each response, and half were not. The reason for asking for justifications was that Medin et al. (2003) had done so, and this design allowed us to examine whether providing justifications would affect responses.

Materials and Procedure

There were seven pairs of inductive arguments, adapted from the nondiversity-by-property-reinforcement items in Medin et al. (2003), as is shown in Table 1. Each pair included an argument based on two taxonomically near animals (such as penguin–eagle) and an argument based on two taxonomically distant animals (such as penguin–polar bear). The taxonomically distant animals nonetheless shared certain salient properties (such as living in a cold habitat for penguin–polar bear). The first five pairs in Table 1 had been validated in terms of similarity judgments collected by Medin et al. (2003). That is, when given a forced choice as to whether the taxonomically near animals or the taxonomically distant animals were more similar, subjects tended to choose the taxonomically near items. The final two pairs produced the opposite pattern of similarity judgments.

Table 1
Stimuli and Results

Taxonomically Near	Inductive Strength (%)	Similarity	Taxonomically Distant	Inductive Strength (%)	Similarity
Penguin–eagle	43.6	3.18	Penguin–polar bear	40.1	4.18
Kangaroo–elephant	50.1	2.16	Kangaroo–frog	49.4	3.02
Camel–rhino	52.9	2.98	Camel–desert rat	39.5	4.18
Polar bear–antelope	52.6	2.16	Polar bear–penguin	40.1	4.18
Chimpanzee–cow	51.2	1.89	Chimpanzee–dolphin	55.9	2.27
Bat–elephant	51.1	1.69	Bat–robin	40.0	4.07
Pig–whale	61.3	1.64	Pig–chicken	48.3	3.53
<i>M</i>	51.8	2.24	<i>M</i>	44.7	3.63

The inductive arguments were given as part of a pen-and-paper survey with 15 questions. The first 2 questions were filler items of a similar form to the main items. The remaining questions were in one of two different random orders and were of the following form:

Given the facts that:

Penguins have Property J.

Eagles have Property J.

How likely is it that:

All animals have Property J?

Each question contained a property referred to by a different letter. Note that whereas Medin et al. (2003) used a variety of conclusion categories, we consistently used *animals*, so as to facilitate the correlational analysis across items. Although the polar bear–penguin argument was used in two different pairs by Medin et al. (2003), we collected data for this argument only once. The subjects were asked to respond to each question on a 0%–100% scale. Half the subjects were also asked to justify each judgment on an additional line.

The similarity condition used a survey with the same 13 pairs as those shown in Table 1, presented in one of two random orders. Corresponding taxonomically near and taxonomically distant pairs did not appear as adjacent items. The instructions were to judge “how similar are each of these pairs of animals” on 1–9 scale, with higher numbers indicating *greater similarity*.

RESULTS

A preliminary analysis did not reveal any influence of justifications in the induction condition. Hence, for subsequent analyses, the data were pooled over the two subconditions. The inductive strength ratings for each argument are shown in Table 1. Overall, there were higher ratings for taxonomically near arguments (51.8%) than for taxonomically distant arguments [44.7%; paired $t(71) = 4.15, p < .001$]. The results were fairly consistent across items. That is, the taxonomically near argument had greater strength than did the taxonomically distant argument for six of seven pairs. Note that Medin et al. (2003) also found higher inductive strength ratings for taxonomically near arguments.

The similarity ratings for each argument are also shown in Table 1. Overall, there were higher similarity ratings for the taxonomically distant arguments (3.63) than for the taxonomically near arguments [2.24; $t(44) = 7.27, p < .001$]. Indeed, each pair of arguments showed this pattern. Note that here, our results depart from those of Medin et al. (2003), who found, for the first five pairs, a tendency to say that the taxonomically near arguments were more similar than the taxonomically distant arguments. Putting together all these results, it appears that there was a diversity effect overall. That is, the taxonomically near arguments were judged as more diverse, and they were judged as being inductively stronger.

The final analysis concerned the correlation between inductive ratings and similarity judgments, taken over the 13 unique arguments. This correlation was $-.86$ ($p < .001$). That is, when the two animals in an inductive argument were judged as more diverse, in terms of having lower similarity ratings, the inductive strength of that argument tended to be higher. Hence, the correlational analysis also showed a diversity effect. Further-

more, the results went in the same direction when correlations within each of the two subsets of arguments were considered ($r = -.62, n.s.$, for the 7 taxonomically near arguments, and $r = -.98, p < .001$, for the 6 taxonomically distant arguments). Although the correlation was not quite statistically significant within one of the subsets, as might be expected when a correlation is computed over a small number of items, these analyses do suggest that the overall pattern is consistent across items.

DISCUSSION

This experiment showed an overall diversity effect, using materials that had been predicted by relevance theory to show nondiversity by property reinforcement. This experiment replicated Medin et al.’s. (2003) results for inductive judgments but showed a different pattern for similarity judgments. Note that the pattern of inductive judgments from Medin et al. (2003) was replicated despite changes in the conclusion categories. These conclusion categories were not presented in the similarity condition, but the different pattern of similarity judgments could possibly reflect other methodological differences. Medin et al. (2003) asked subjects to make forced-choice judgments based on “general similarity, not on associations” and found that taxonomically near animals (e.g., penguin–eagle) tended to be chosen over taxonomically distant animals (e.g., penguin–polar bear). It is possible that the forced-choice method, or telling subjects to ignore associations, could have focused attention on the biological characteristics of the animals, rather than on their behavior or living environment. Indeed, it could be argued that telling subjects to ignore associations is desirable, because it might facilitate a clear-cut assessment of similarity without undue attention to shared properties, such as living in a cold habitat. In the extreme, similarity judgments themselves might be affected by a *property-reinforcement* effect.

Therefore, in a posttest, we asked 68 additional subjects from the same population to make forced-choice similarity judgments on the seven taxonomically near versus taxonomically distant comparisons. There were 35 subjects in the basic condition, in which people were asked simply to choose which pair of categories was most similar. There were 33 subjects in the enhanced condition, which had additional instructions to “base your judgments on general similarity, and not on associations.” The materials were presented in two random orders. In the basic condition, the taxonomically distant pairs were chosen 71.0% ($SE = 3.23\%$) of the time, significantly different from chance [$t(34) = 6.20, p < .001$]. Likewise, in the enhanced condition, the taxonomically distant pairs were chosen 65.4% ($SE = 3.73\%$) of the time [$t(32) = 4.12, p < .001$]. Although the proportion of taxonomically distant choices was slightly reduced in the enhanced condition, the difference between conditions was not statistically significant [$t(66) = 1.15$]. The two conditions were pooled together for an item analysis.

For six of the seven comparisons, the proportion of taxonomically distant choices was greater than 50%. In five of these six comparisons, the proportion was significantly different from chance, using a binomial test ($p < .05$). There was just one comparison (chimpanzee–cow versus chimpanzee–dolphin) for which the proportion of taxonomically distant choices (32%) was significantly less than 50%.

In sum, the forced-choice similarity judgments in our posttest were consistent with the individual similarity ratings in the experiment. Taxonomically distant pairs were generally judged as more similar (with one exception), and because these had also been judged as inductively weaker, there was a diversity effect overall. Our forced-choice results were not consistent with those reported by Medin et al. (2003). Although our task was comparable to that in Medin et al. (2003), and the sample size was larger (68 in total, as compared with 20), it is possible that other methodological differences could have led to the different results, such as the different subject populations (Britain vs. Belgium) and the fact that Medin et al.'s (2003) similarity test contained 14 additional questions not used in our own posttest. Although we acknowledge that similarity judgments are context dependent (Medin, Goldstone, & Gentner, 1993), we did not find much effect of using a forced-choice, rather than a similarity-rating, procedure, nor was there much effect of telling people to ignore associations.

Further Experimental Evidence

In two further experiments, we (Heit et al., 2005) have examined another stimulus set for which relevance theory would predict nondiversity by property reinforcement. These stimuli were adapted from Heit and Rubinstein (1994), who also had created taxonomically near pairs of animals (such as whales and bears) and taxonomically distant pairs of animals (such as whales and tunas). Heit and Rubinstein had collected two kinds of similarity ratings for these stimuli; similarity with respect to anatomy and similarity with respect to behavior. The taxonomically distant animals were overall considered less similar in terms of anatomy than the taxonomically near animals. However, the taxonomically distant animals were considered more similar in terms of behavior than the taxonomically near animals were. In other words, the taxonomically distant pairs were considered more diverse overall, but they still shared some salient properties.

The two further experiments followed the same procedure as that for the induction condition of the main experiment reported here. The test questions in the first experiment were of the following form:

Given the facts that:

Whales have property C.

Bears have property C.

How likely is it that:

All animals have property C?

The second experiment emphasized that the property in question referred to an animal's behavior, in case this would encourage a focus on other behavioral features and, possibly, make a nondiversity-by-property-reinforcement effect more likely. Nonetheless, both experiments showed diversity effects, rather than nondiversity effects. In general, greater diversity in terms of anatomical similarity led to stronger inductive judgments. For example, the correlation between anatomical similarity ratings and inductive strength judgments was $-.85$ in the first experiment and $-.64$ in the second experiment. (There was approximately a zero correlation between behavioral similarity and inductive strength.) Hence, in three experiments for which relevance theory had predicted nondiversity by property reinforcement, we have instead found diversity effects.

Reassessing Relevance Theory

The theoretical implication of the results is that these diversity effects can be explained by previous models of inductive reasoning (Heit, 1998; Osherson et al., 1990; Sloman, 1993). Hence, the results do not serve as distinctive evidence for Medin et al.'s (2003) relevance theory over other accounts.

However, Medin et al. (2003) presented seven other novel phenomena that would also serve to distinguish relevance theory from previous models of inductive reasoning. For example, in the causal nondiversity phenomenon, causal knowledge is used to override the usual diversity effect, as is illustrated by the following arguments:

(5) Sparrows have property X.

Dogs have property X.

All living things have property X.

(6) Sparrows have property Y.

Seeds have property Y.

All living things have property Y.

In this example, sparrows and seeds are judged more diverse than sparrows and dogs, yet Argument 5 is judged stronger than Argument 6, because people consider the causal mechanism for transferring properties from seeds to sparrows and judge that this mechanism would not apply generally to all living things. This kind of causal reasoning could not be explained by previous models. More generally, Medin et al. (2003) reported a substantial body of evidence for which relevance theory would provide a distinctive explanation.

In conclusion, although the relevance theory of category-based induction appears to predict when other sources of background knowledge will overcome diversity, the evidence presented here suggests that, in the absence of such information, the diversity effect is robust.

REFERENCES

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

- HAYES, B. K., GOODHEW, A., HEIT, E., & GILLAN, J. (2003). The role of diverse instruction in conceptual change. *Journal of Experimental Child Psychology*, **86**, 253-276.
- 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., & HAHN, U. (2001). Diversity-based reasoning in children. *Cognitive Psychology*, **43**, 243-273.
- HEIT, E., HAHN, U., & FEENEY, A. (2005). Defending diversity. In W.-K. Ahn, R. L. Goldstone, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside of the laboratory: Essays in honor of Douglas L. Medin* (pp. 87-99). Washington, DC: American Psychological Association.
- HEIT, E., & RUBINSTEIN, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 411-422.
- HOMA, D., & VOSBURGH, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 322-330.
- KIM, N. S., & KEIL, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory & Cognition*, **31**, 155-165.
- KINCANNON, A., & SPELLMAN, B. A. (2003). The use of category and similarity information in limiting hypotheses. *Memory & Cognition*, **31**, 114-132.
- LÓPEZ, A. (1995). The diversity principle in the testing of arguments. *Memory & Cognition*, **23**, 374-382.
- LÓPEZ, A., ATRAN, S., COLEY, J. D., MEDIN, D. L., & SMITH, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, **32**, 251-295.
- MEDIN, D. L., COLEY, J. D., STORMS, G., & HAYES, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, **10**, 517-532.
- MEDIN, D. L., GOLDSTONE, R. L., & GENTNER, D. (1993). Respects for similarity. *Psychological Review*, **100**, 254-278.
- OSHERSON, D. N., SMITH, E. E., WILKIE, O., LÓPEZ, A., & SHAFIR, E. (1990). Category-based induction. *Psychological Review*, **97**, 185-200.
- POSNER, M. I., & KEELE, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.
- PROFFITT, J. B., COLEY, J. D., & MEDIN, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 811-828.
- SALMON, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.
- SLOMAN, S. A. (1993). Feature-based induction. *Cognitive Psychology*, **25**, 231-280.
- TENENBAUM, J. B., & GRIFFITHS, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral & Brain Sciences*, **24**, 629-641.

(Manuscript received December 24, 2003;
revision accepted for publication June 7, 2004.)