Memory & Cognition
2004, 32 (2), 235-243

Category-based induction: An effect of conclusion typicality

JAMES A. HAMPTON and IBEN CANNON
City University, London, England

Category-based induction involves the willingness of a thinker to project some newly learned property of one or more classes of objects to another class on the basis of their shared membership in a common superordinate category. Previous research has established that the perceived strength of arguments of the form "Class A has Property P; therefore, Class B has Property P" is influenced by the similarity of A to B and by the typicality or representativeness of A in a shared category, superordinate to both A and B. (The nature of P is also crucial, but we do not examine it in this study.) There is, however, no prior evidence that the relation between B and the category is influential. Three experiments were designed to test whether the typicality of B in the superordinate category also has an effect on inductive argument strength. By using multiple regression (Experiment 1) and an experimental design (Experiment 3), an effect of conclusion typicality was found, so that people are more willing to project properties to more typical conclusions. Experiment 2 ruled out conclusion familiarity as a potential confounding variable. The results are interpreted in the light of current models of category-based induction.

In a pioneering study, Rips (1975) posed problems of the following form to students:

> If all *rabbits* on a particular island have a new type of contagious disease, then what proportion of *mice* would be expected to have the disease?

This type of argument has been termed inductive in that it is presumably through this form of reasoning that more general hypotheses are generated on the basis of individual observations. In particular, in contrast to deductive forms of reasoning, such as syllogisms or conditionals, there is no clear way of determining what the correct answer in such a case may be. It is, therefore, of great interest to determine the factors within such arguments that influence people when they come to make a judgment of argument strength. This psychological question is, of course, largely independent of the equally interesting epistemological question concerning the grounds that would *actually* justify confidence in such an argument.

Rips (1975) worked with biological categories, such as *birds.* First, he took pairwise similarity judgments for a set of words from the same category, such as a sample of different bird names, together with the category name *bird* itself. These similarity judgments were then scaled with multidimensional scaling (MDS) to generate a two-dimensional similarity space. Effectively, each concept name was placed on a two-dimensional map, with the constraint that proximity on the map corresponded as closely as possible to mean rated similarity for any pair of concepts. A second group of participants was given a set of arguments of the form cited above, each argument created by taking a different pair of concepts from the set of category members and arranging them in a particular order. This new group made judgments of the strength of the inductive arguments. A regression was then used in the analysis to try to predict the rated strength of an argument in terms of the relative positions of the two concepts in the similarity space. For an argument of the form "If all A have a particular property, then what proportion of B will have it?" A is the premise term and B the conclusion term, and let us call C the most specific superordinate category that contains both A and B (but is not mentioned explicitly in the argument). In Rips's study, there were, then, three possible predictors of the argument strength (see Figure 1): (1) the distance of Premise A from Category C, (2) the distance of Premise A from Conclusion B, and (3) the distance of Conclusion B from Category C.

When these three variables were entered into the regression, Rips (1975) found that only the first two predictors were needed to account for variations in argument strength. The closer the premise was to the category (i.e., the more typical it was) and the closer the premise was to the conclusion (i.e., the more similar the premise and the conclusion), the stronger was the argument. The typicality of Conclusion B (the proximity of B to C) was not a significant predictor of strength. As a logical corollary of this result, Rips also showed that there is a premise–conclusion asymmetry effect in regard to typicality. When one of the

The authors acknowledge Neville Austin, Evan Heit, Steven Sloman, and members of the London Concepts Group and the City University Memory and Cognition Group for helpful discussions concerning the research and Daniel Osherson and two anonymous referees for their suggestions on the manuscript. Special thanks are due Claire Simmons for help with data collection and general support. Correspondence concerning this article should be addressed to J. A. Hampton, Psychology Department, City University, Northampton Square, London EC1V OHB, England (e-mail: hampton@city.ac.uk).

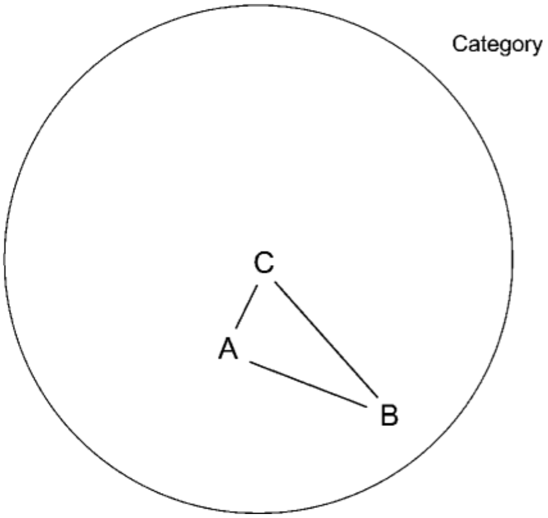

Figure 1. Representation of a category prototype (C), with a typical instance (A) and an atypical instance (B). As the distance AC gets shorter, the correlation between AB and BC, as B takes different positions, becomes higher. AB represents premise–conclusion similarity, whereas AC and BC represent the premise and the conclusion typicalities, respectively.

category members is more typical than the other, the strength of the argument Typical → Atypical is stronger than the reverse argument. This asymmetry follows automatically from the finding that premise typicality affects argument strength, whereas conclusion typicality does not, since clearly, the argument pair Typical → Atypical will have a more typical premise than will the reversed pair Atypical → Typical. (For the sake of argument, it is assumed that the similarity between the A and B terms is the same measured in each direction, as will necessarily be the case when proximity measures are taken from MDS solutions. Evidence for this assumption has been provided by Aguilar & Medin, 1999.)

Subsequent work by Osherson, Smith, Wilkie, Lopez, and Shafir (1990) laid out evidence for a set of 13 phenomena relating to category-based induction, using categories at different levels and exploring arguments with either single or multiple premises. One methodological advance on Rips (1975) was their introduction of an improved format for the task. Asking about disease prevalence may bring to mind all kinds of different properties of the animals, such as their diet and habitat, that could influence the answer (see, e.g., Coley, Medin, Proffitt, Lynch, & Atran, 1999). In order to keep the logical structure of the argument free from any specific effects of stored knowledge or background theory, other than knowledge of the taxonomic hierarchy of creatures categories, Osherson et al. adopted the following format:

> Rabbits use serotonin as a neurotransmitter
>
> therefore
>
> Mice use serotonin as a neurotransmitter.

Osherson et al. (1990) also tested for premise–conclusion asymmetry, but with equivocal results. For example, when asked to choose directly between the two arguments Bat → Mouse versus Mouse → Bat, responses were evenly divided (a mouse would be a typical mammal, and a bat an atypical mammal). Likewise, when the participants, rather than making a forced choice, gave ratings of the conditional probability of the conclusion given the premise, there was no significant asymmetry effect. The only condition in which a reliable asymmetry effect *was* obtained was a condition in which (1) the critical forced choice between the two arguments was placed in a context with other filler pairs and (2) instructions were given that "there is always a difference in how much reason the facts of an argument give to believe its conclusion—however small this difference may be, we would like you to indicate for which argument the facts provide a better reason to believe the conclusion." Having noted this restriction on the generality of the asymmetry effect, one should also note that Osherson et al. did not state how many pairs of such arguments were tested per experiment, so their results (or lack of them) were quite possibly due to item-specific factors and to a small sample size of items. It appears, then, that the evidence for premise–conclusion asymmetry may be fairly weak. This weakness could reflect either an absence of a premise typicality effect or the presence of an equally strong conclusion typicality effect.

The lack of an effect of conclusion typicality reported by Rips (1975) has been further confirmed in several developmental studies, in some of which adult groups were also used as controls (Carey, 1985; Gelman & O'Reilly, 1988). In an extensive review of the literature, Heit (2000) concluded "there have been no reports to date of independent effects of the typicality of the conclusion category as opposed to the premise category" (p. 576). Perhaps the only result that suggests that there may be such an effect comes from a study by Sloman (1998), in which a category name was taken as the *premise* of the argument. Arguments of the form Plants → Mosses should be perfectly convincing, provided that the participant agrees with the premise (such as *all plants have bryophytes*) and agrees that all mosses are indeed plants. However, Sloman showed that people still felt more confident in arguments that led to more typical conclusions—for example preferring Plants → Flowers to Plants → Mosses. Sloman's (1998) result, therefore, sits rather uneasily with the apparent lack of a conclusion typicality effect when the premise is a category member, rather than a category name. We therefore decided to investigate the issue of conclusion typicality further.

The aim of the present research was to set up an experimental test of the existence of a conclusion typicality effect. To this end, we considered carefully the best way in which to select materials for the test. It was not sufficient simply to select a premise and then two conclusions of differing typicality, all from the same category, since if typical items have a tendency to be more similar to one another than do typical to atypical items, an effect of conclusion typicality could simply reflect a difference in premise–conclusion similarity.

Sets of category members were, therefore, needed that were well balanced for premise typicality and premise–conclusion similarity but would vary maximally on conclusion typicality. One of the problems with earlier research is that if the premise concept is a *typical* member of the category, it follows that there will be a close correspondence between premise–conclusion similarity and conclusion typicality. As the points A and C in Figure 1 draw closer together, so the lengths of AB and CB are more and more constrained to be very similar. Given that premise–conclusion similarity (AB) is known to have the strongest effect on inductive argument strength, it is unlikely that there would be an independently detectable effect of conclusion typicality simply because of the multicollinearity arising from the use of a very typical premise. Accordingly, we chose premises of intermediate typicality, which would enable us to choose conclusions of varied typicality but matched similarity to the premise.

EXPERIMENT 1

The first experiment adopted Rips's (1975) regression method. Independent measures were obtained of premise typicality, conclusion typicality, premise–conclusion similarity, and the inductive strength of the argument from premise to conclusion. Inductive strength was then regressed on the other three variables.

Method

Participants. Forty-eight students at City University, London participated on a voluntary basis to provide the initial set of normative ratings, 25 for typicality and 23 for pairwise similarities. An additional 19 students volunteered to provide a second set of similarity ratings for new pairings of the original items. The final balanced set of materials was then given to a third group of 36 students, who rated the inductive strength of the premise–conclusion arguments, participating in return for course credit. There was no overlap between any of the groups.

Materials. Materials were generated in triplets within three biological categories: birds, mammals, and insects. Pretesting involved first obtaining ratings of typicality for a larger set of items in their respective categories and obtaining similarity ratings for paired premise and conclusion concepts in each category. Order within the list was randomized within categories, except that, for similarity ratings, pairs with the same premise (e.g., *horse–cow* and *horse–bison)* were kept maximally apart in the list. Order of premise and conclusion concepts within a pair was counterbalanced across materials. Each category was presented on a separate page of a three-page booklet, and page order was balanced across participants for both tasks.

Analysis of the first set of ratings revealed that the initial construction of triplets had not succeeded in balancing the similarity of premise–conclusion pairs within each triplet. The same items were, therefore, re-paired into new triplets, and a second set of similarity ratings was obtained. It was then possible to construct a total of 22 triplets, 8 for mammals and 7 each for birds and insects.

Procedure. For the ratings of typicality and similarity, the participants were tested in a classroom setting and completed the booklets in their own time without conferring. The typicality booklet contained three pages, one for each category, and used a 1–7 numbered scale, with 1 meaning *very typical* and 7 *very atypical*. Instructions emphasized the difference between familiarity and typicality (see Hampton & Gardiner, 1983). Similarity ratings were also blocked within category, in order to ensure that the basis used for similarity was relevant to the task. The category name was printed at the top of the page for both tasks. Similarity instructions asked the participants to "rate each pair according to how similar or dissimilar the two instances appear to you within the category to which the pair belongs." A 1–7 numbered scale was again used, with 1 representing a *very similar pair* and 7 a *very dissimilar pair*. In both tasks, if a participant did not know any of the items, he or she was instructed to underline the item and leave the scale blank. Seventeen participants (25%) were excluded from the analysis because either they marked more than 20% of the items as unknown or they failed to engage with the task (e.g., marking every item as *very typical*).

From an initial set of 30, a final set of 22 triplets was selected after re-pairing and remeasuring similarities. Within each triplet of Premise A, Typical Conclusion B_{typ}, and Atypical Conclusion B_{atyp}, the similarity of A to each conclusion was matched, and the conclusions differed maximally in their typicality in the category. All the items in the final set were known to at least 75% of the participants. (Some degree of unfamiliarity for the atypical items was necessary in order to generate triplets that were otherwise matched on similarity. Experiment 2 returned to the question of whether familiarity might be confounding the effects of typicality.) A list of the triplets used, together with their mean ratings, is shown in Table 1.

Finally, the 22 triplets were used to construct 44 inductive arguments, taking the premise with either of the two conclusion terms. As in Osherson et al. (1990), 44 different biologically plausible blank predicates were used, such as "needs Vitamin K for liver function." Each argument was quantified with the word *all* in order to stress that the premise and the conclusion applied to whole kinds, rather than to individuals. One argument from each triplet was arranged in a random order in one booklet, whereas the other was placed in a second booklet. Order was random but blocked by category. Each booklet contained equal numbers of typical and atypical arguments. For each category, two additional filler items were included in the first and sixth positions in the list, in order to disguise the design. Two additional booklets were constructed by reversing the order of arguments within each category list (while keeping the filler items in the same place). The order of the three categories in the booklets was also counterbalanced across participants. The booklets were distributed to students in a classroom setting and were completed without time constraint or conferring. Instructions were to consider the first statement as being a true fact and then to judge the probability of the conclusion's being true in the light of the evidence provided by the fact. Ratings were made on a 10-point scale, with 1 representing *no faith in the argument's being true* (*very unlikely*) and 10 representing *strong belief that the argument could be true* (*very likely*). As previously, the participants were told to underline any items they did not know and to leave the scale blank.

Results

Reliability of ratings. Across the 22 triplets, mean rated similarity for premise–typical-conclusion and for premise–atypical-conclusion pairs was identical at 4.43 (*SD*s = 0.89 and 1.04, respectively). The average standard error for individual item pair similarity ratings was 0.37. Mean rated typicality was 2.30 (*SD* = 0.44) for the premise items, 1.72 (*SD* = 0.36) for the typical-conclusion items, and 3.41 (*SD* = 0.61) for the atypical-conclusion items. Across all premises, the variance in typicality was 0.19, whereas across all conclusions, the variance in typicality was 0.98. There was, therefore, five times as much

Table 1
Triplets Used in Both Experiments, Together With Mean Ratings for Premise Typicality, Similarity Between the Premise and Each Conclusion, and Typicality and Familiarity (From Experiment 2) for the Conclusion Items, Also With Argument Strength (Experiment 1)

Premise	Mean Premise Typicality	Mean Premise Familiarity	Typical Conclusion/ Atypical Conclusion	Mean Premise–Conclusion Similarity	Mean Conclusion Typicality	Mean Conclusion Familiarity	Mean Argument Strength
Category = Mammals							
Koala	2.44	3.39	tiger	5.71	1.81	4.28	4.28
			guinea pig	5.47	3.13	4.28	2.78
Dog	1.25	4.89	fox	2.53	1.81	4.61	6.00
			coyote	2.00	3.87	2.78	4.83
Whale	2.56	4.00	grizzly bear	6.06	2.19	3.56	3.78
			bat	6.31	3.56	3.94	2.78
Horse	1.75	4.72	cow	3.76	1.44	4.67	5.89
			bison	3.88	2.69	1.89	5.25
Boar	2.80	2.67	deer	4.71	1.81	4.06	4.72
			walrus	5.47	3.71	2.83	4.64
Zebra	2.31	3.89	hippopotamus	5.59	2.13	3.78	3.28
			squirrel	5.71	2.94	4.50	3.28
Hare	2.56	3.56	goat	5.00	1.81	4.39	3.39
			rat	4.35	3.06	4.50	4.06
Wolf	2.13	3.83	hyena	3.00	2.63	3.67	5.89
			dingo	2.65	4.00	2.33	6.56
Category = Birds							
Vulture	2.40	3.11	sparrow	4.94	1.06	4.28	6.06
			quail	4.69	3.54	2.72	4.62
Pelican	3.00	3.50	parrot	3.18	1.38	4.17	6.47
			toucan	3.13	3.42	2.56	5.19
Heron	2.31	2.94	pheasant	4.50	2.19	3.61	5.07
			emu	4.07	4.14	3.06	4.56
Eagle	1.31	3.94	crow	3.82	1.19	4.17	5.28
			cockatoo	5.00	2.38	3.33	4.89
Stork	2.93	3.33	falcon	3.44	1.94	3.06	6.12
			goose	3.19	3.00	3.94	6.28
Chicken	2.38	4.61	magpie	4.81	1.57	3.61	5.17
			flamingo	4.82	3.00	3.72	5.67
Duck	2.00	4.56	canary	4.50	1.44	3.61	4.89
			ostrich	4.12	3.33	3.50	6.83
Category = Insects							
Centipede	2.47	3.61	beetle	4.35	1.73	4.33	5.81
			silverfish	4.57	5.08	2.94	3.88
Butterfly	2.69	4.50	ant	5.35	1.44	4.94	4.22
			tick	5.50	3.29	2.89	3.35
Daddy long legs	2.19	4.06	cockroach	4.41	1.63	4.06	4.56
			praying mantis	3.85	4.33	2.11	4.00
Dragonfly	2.38	3.72	flea	4.53	1.94	3.39	3.83
			maggot	5.06	3.53	3.28	3.56
Ladybird	2.13	4.28	spider	4.65	1.50	4.78	3.94
			termite	4.44	2.88	3.00	4.06
Wasp	2.06	4.39	cricket	4.71	1.75	3.67	3.81
			scorpion	4.94	3.19	3.50	4.11
Locust	2.53	3.06	mosquito	3.94	1.50	4.11	4.39
			earwig	4.23	3.00	2.94	5.14

variance in the conclusion typicalities as in the premise typicalities. As had been intended, the materials maximized the manipulation of conclusion typicality. A consequence of this procedure, however, was to produce greatly truncated variance in premise typicality. Reliability of the premise–conclusion similarity judgments across the 44 arguments was 0.84 (Spearman–Brown). Across the 66 typicality judgments, reliability was 0.86. This broke down into 0.89 for the 44 conclusion typicalities, but only 0.52 for the 22 premise typicalities (because of the truncated variance).

Mean inductive strength was calculated for each of the 44 arguments and is shown in Table 1. Thirty-eight data points (fewer than 5%) were treated as missing because of unfamiliarity or failure to complete a rating. No participants or triplets needed to be omitted because of undue levels of missing data. Corrected split-half reliability for inductive strength was 0.82.

Regression analysis. All 44 arguments were entered into a multiple regression analysis to predict mean rated argument strength on the basis of the three predictors: premise–conclusion similarity, premise typicality, and conclusion typicality. (Because of the direction of scoring of the scales, all the predictors were expected to enter with negative coefficients.) As was expected, the regression showed a strong effect of premise–conclusion similarity ($\beta = -.733, p < .001$). There was no reliable premise typicality effect ($\beta = .112, p > .10$), probably because of the deliberately reduced variance and, consequently, low reliability for this measure. There was, however, a significant effect of conclusion typicality ($\beta = -.246, p = .029$). Adjusted R^2 was .50. When dummy variables were included to take account of the factor of category, the conclusion typicality effect was somewhat smaller but remained significant ($\beta = .195, p = .047$). Further tests showed that after entering premise–conclusion similarity as a first step, conclusion typicality would enter next with a significant β ($p = .035$), whereas premise typicality would not ($p = .46$).

Lorch and Myers (1990) have suggested that a more valid way to analyze regression data when using items as the random variable is to perform separate regression analyses on each individual participant and then to combine the results. The set of ratings given by each of the 18 participants who rated inductive strength was taken individually, and 18 regression analyses were run to predict their individual ratings on the basis of the three predictors. Mean multiple R was .40, and mean R^2 was .18. The results confirmed a strong effect of premise–conclusion similarity [mean $\beta = -.34$; one-sample $t(17) = -8.00$, $p < .001$] and a significant effect of conclusion typicality [mean $\beta = -.12$; one-sample $t(17) = -3.27, p < .005$]. As before, there was no significant effect of premise typicality. The β coefficient was in the predicted direction for 17 out of 18 participants for the similarity effect and for 14 out of 18 participants for the typicality conclusion effect, 5 of which were individually significant at .05, one-tailed.

Discussion

The aim of Experiment 1 was to select a set of materials in such a way as to optimize the chances of obtaining an effect of conclusion typicality on the judged strength of a simple inductive argument. The normal correlation between premise–conclusion similarity and conclusion typicality was prevented from confounding the test of the hypothesis by deliberately choosing premises of intermediate typicality. With this correlation held artificially low ($r = -.07, p > .5$), variations in conclusion typicality were indeed found to predict inductive strength. More specifically, the higher the typicality of the conclusion concept in an inductive argument, the more strongly was the argument rated.

EXPERIMENT 2

One issue that remains a potential problem for interpreting the results of Experiment 1 concerns the differences in familiarity of the typical and the atypical materials within the triplets. In order to construct well-balanced triplets, it proved necessary to allow atypical conclusion concepts to be prima facie less familiar than typical conclusion concepts. Not all atypical items were unfamiliar—for example, *rat* and *squirrel* are probably more familiar than *hyena* or *grizzly bear* to a British sample. However, should it turn out that familiarity of conclusion concepts predicts the strength of an inductive inference, our interpretation of the conclusion typicality effect is at risk. Indeed, some unpublished data from Collister and B. Tversky (personal communication) suggests that people may prefer arguments with less familiar premises and more familiar conclusion terms.

In Experiment 2, therefore, a further sample of students from the same population was asked to provide ratings of familiarity for all 66 concepts used in Experiment 1. Mean familiarity ratings were calculated and correlated with the predictors and dependent variable from that experiment.

Method

Participants. Eighteen students at City University, London participated for course credit. None took part in any of the other experiments.

Procedure. Booklets were created in which the concepts from the triplets were listed in alphabetic order, blocked by category. A 5-point scale was provided, and the participants rated the familiarity of each concept from 1 = *very unfamiliar* to 5 = *very familiar*.

Results

Mean familiarity ratings were calculated for each of the 66 concepts. Split-half reliability was estimated at 0.92. Mean familiarity across concepts ranged from 4.6 and above for highly familiar concepts such as *dog*, *duck*, and *ant*, down to 1.9 for the less familiar *bison*. Mean familiarity was 3.8 ($SD = 0.6$) for the premises, 4.0 ($SD = 0.5$) for the typical-conclusion concepts, and 3.2 ($SD = 0.7$) for the atypical conclusions. Within the triplets, 16 of the 22 had a more familiar typical- than atypical-conclusion term. However, when compared with the results of Experiment 1, familiarity had *no correlation* with inductive strength of arguments. Familiarity of the premise term correlated with inductive strength, with $r = .03$, and familiarity of the conclusion term correlated with strength, with an r of $-.07$. Furthermore, the familiarity variables failed to enter any of the regression models examined in Experiment 1 with significant coefficients.

There was, therefore, no evidence that the observed significant effect of conclusion typicality was owing to differences in familiarity.

EXPERIMENT 3

To provide a further test of generality for the conclusion typicality effect, Experiment 3 adopted a quasi-experimental design, so that generalization across both items and subjects could be tested. Instead of making ratings of strength for individual arguments, a forced-choice procedure was used in which a pair of arguments was presented and the participants were asked to select the stronger. Osherson et al. (1990) used a similar procedure. The forced-choice procedure helps to focus attention on the key factor of the conclusion item, since the arguments in a pair had the same premise and the same predicate but differed in their conclusions.

Method

Participants. Forty-eight students at City University, London volunteered to participate, either with no incentive or for course credit. None of the participants had taken part in any of the earlier experiments.

Materials. The same set of 22 triplets was used as in Experiment 1. The triplets were organized into pairs with a common premise and two different conclusions, one typical and one atypical. The same filler items were used within each category list to disguise the manipulation of conclusion typicality.

Procedure. Booklets contained 28 argument pairs, comprising the 22 critical test pairs blocked by category and the 6 filler pairs. The participants completed the booklets in a group setting, without time constraint. An example of an argument pair was as follows:

> a) All Vultures have sesamoid bones, therefore all Robins have sesamoid bones.
>
> b) All Vultures have sesamoid bones, therefore all Ostriches have sesamoid bones.

One argument was presented on the left of the page, and the other argument opposite it on the right of the page. Across booklets, typical and atypical arguments occurred equally often on the left or the right. The participants had to indicate which of the two they considered a stronger argument—that is, which they felt was more likely to be true. Order of critical pairs was randomized within categories, with the filler arguments in Positions 1 and 6. A second set of booklets used the reverse random order for critical pairs. The order of categories within the booklets was also counterbalanced, leading to 12 different booklets. Four participants completed each booklet. As before, if a concept was unknown, the participants were asked to underline it and move on to the next pair.

Results

Less than 6% of the data were omitted because of unknown items, and the data from all the participants were used. The argument with the more typical conclusion was selected 58% of the time, where chance would have been 50%. This result was significant across participants [$t(47) = 3.89, p < .001$, two-tailed]. Thirty-four out of 48 participants showed the effect. Similar t tests were significant for each of the three categories individually, with p values between .05 and .003. Across items, 17 out of 22 paired arguments had a greater proportion of participants selecting the more typical conclusion ($p < .01$ on a sign test). This result was also significant on a one-sample t test across items [$t(21) = 2.37, p < .027$, two-tailed]. There was no significant correlation between the proportion of participants choosing the more typical conclusion for an item and the difference in the rated familiarity from Experiment 2 of the two conclusions ($r = .1$). There was, therefore, again no evidence that familiarity plays a role in judging inductive strength.

Discussion

Experiment 3 confirmed the finding from Experiment 1 that the typicality of a conclusion concept can affect the perceived strength of an inductive argument. When given a choice between two arguments, one with a typical conclusion and one with an atypical conclusion (with respect to the common superordinate category), the participants showed a significant tendency to prefer the typical one. The effect was relatively small (as in Experiment 1) but was consistent, with 71% of the participants and 77% of the items showing the predicted pattern.

GENERAL DISCUSSION

The results presented here provide a clear demonstration of a robust effect of the typicality of the conclusion term in category-based inductions. Before discussing our own account of the data, we next will discount three alternative accounts of the data and consider the implications for Sloman's (1993) and Osherson et al.'s (1990) models.

Choice of Appropriate Superordinate

If Osherson et al.'s (1990) model is correct, it could account for our data by supposing that premises and typical conclusions were members of more narrowly defined superordinates than the corresponding premises and atypical conclusions. For example [*duck* + *canary*] may cue retrieval of *flying birds*, whereas [*duck* + *ostrich*] would cue retrieval of *birds*. Similarly [*horse* + *cow*] might cue *farm animal*, whereas [*horse* + *bison*] would cue just *animal*. The more narrowly defined the superordinate category, the more likely it is that the premise term will provide fuller coverage of the category, and hence, the stronger will be the argument.

Basis of Similarity

A related notion would be that the use of the biological blank predicate (e.g., *uses serotonin as a neurotransmitter*) could have triggered a different similarity metric from that considered when people just rated similarity alone (see Heit & Rubinstein, 1994). For example, about half of the triplets have an atypical conclusion term that comes from a different geographical region from the other two terms (e.g., *eagle–crow–cockatoo*). Perhaps people might have used geographical region as an indicator of deeper biological similarity for the inductive judgment but ignored it when making general similarity judgments in the pretest.

Although these are both plausible explanations, there was no support for them in the data. To illustrate this claim, Table 1 lists the triplets within each category in descending order of their effect size (argument strength for the typical conclusion minus argument strength for the atypical conclusion) in Experiment 1. The strongest effects of conclusion typicality in Experiment 1 were in triplets for which neither of these accounts apply (e.g., *centipede–beetle–silverfish* or *koala–tiger–guinea pig*). In addition, the triplet that went most strongly against the hypothesis was one to which both arguments *do* apply (*duck–canary–ostrich*). *Ostrich* is from a different evolutionary niche and is not in the *flying bird* category, but (contrary to the overall trend) the argument from *duck* to *ostrich* was rated as stronger than that from *duck* to *canary*. When coded as binary variables (using the first author's intuitions), neither of these hypotheses entered into the regressions for Experiment 1 with significant coefficients, whereas the effect of conclusion typicality remained strong and significant. Of course, a fully adequate test of these accounts would require further work, predicting argument strength from two new data sets. One would be data on the closest common superordinate that people associate with a pair of items, and the other would be similarity judgments taken from a biological perspective.

Nonspecific Effects of Typicality

Our results depend critically on the preference of the participants for arguments that have more typical conclusion terms. An interesting third possibility is that there may be some overall nonspecific bias toward accepting the truth of statements about typical concepts, regardless of any other considerations. The test for this would be to obtain estimates of the truth of the argument conclusions in the absence of the premises. It is implausible to expect such a bias, given that it would be reflecting typicality in relation to a category that is not mentioned in the task. Furthermore, if there were a bias, it would more plausibly relate to familiarity and work the other way. The more ignorant we are of an item, then perhaps the more willing we are to accept that any given predicate should be true of it. We, therefore, consider this account unlikely.

Given the demonstration that the typicality of the conclusion concept may affect inductive strength, our discussion now will turn to how two particular current models may or may not be able to accommodate such an effect.

Sloman's Feature-Based Model

The first model to be considered was proposed by Sloman (1993). In his feature-based model, when premises are introduced, the features possessed by those premise concepts are associated to the predicate. The strength of the argument associating the predicate to the conclusion is then determined by the feature overlap between the conclusion term and the set of activated features, divided by a measure of the "magnitude" of the conclusion term. More specifically, argument strength will increase with premise–conclusion similarity but will decrease the greater the number of features possessed by the conclusion term. Sloman's (1993) model is entirely feature based and requires no consideration of set inclusion relations. It is, therefore, well suited to the approach proposed above. More radically, however, it does not require activation of any superordinate category in the case of single-premise arguments, such as those considered in this article. All typicality effects are handled either by feature overlap or by the degree of richness of the conclusion concept term. For example, to account for the possible asymmetry between Typical → Atypical and Atypical → Typical arguments, it is proposed that more typical conclusions will tend to have richer feature representations and, hence, will have weaker argument strengths. Sloman (1993) discussed evidence based on six pairs of items for which typicality and richness of representation could be disconfounded. He concluded that the number of features possessed by the conclusion term is the critical factor in making weaker arguments (see also Sloman & Wisniewski, 1992).

If this is the correct account of the asymmetry effect and if typical concepts are distinguished by having richer feature representations, holding the premise constant (as was done here) and varying typicality of the conclusion should show an *inverse* conclusion typicality effect; more typical conclusions should show weaker strength ratings, to the extent that they have richer feature representations. To explain the advantage for arguments with typical conclusions, Sloman would have to argue (in contradiction to his account of the asymmetry effect) that the typical conclusion items used in these studies have *less rich* feature representations. The reader can check the plausibility of this idea in Table 1, where the triplets are listed in descending order of effect size. To take the two triplets with the strongest effect, it would imply that *beetle* has a less rich representation than *silverfish* and that *tiger* has fewer features represented than *guinea pig*.

Note also that Experiment 2 failed to show any effects of familiarity of the premise and conclusion concepts on inductive argument strength, whereas one might reasonably expect concepts with richer feature representations to be judged as more familiar.

Osherson et al. (1990)

The second model is Osherson et al.'s (1990) coverage model, which suggests that inductive strength in single-premise arguments is determined by two things: the similarity between premise and conclusion terms and the average similarity between the premise and other members of the lowest superordinate category that includes both the premise and the conclusion. Their model, therefore, predicts that if premise typicality and premise–conclusion similarity are held constant, there will be no conclusion typicality effect. Although their model is clearly successful at accounting for a wide range of phenomena, nonetheless our data suggest that it is in need of modification.

We suggest, like Osherson et al. (1990), that there are two *routes* to argument strength: one direct route through the similarity between the mental representations of the premise and the conclusion concepts and a second indirect route via category membership. The first route, via similarity, is presumably a very general route, applying to any pair of items regardless of their category membership. It is the second route, involving membership in the common superordinate category, that needs revision. The indirect route involves two stages, both probably driven by a belief that members of biological categories share hidden biological properties. One simple modification of Osherson et al.'s model would propose that when the superordinate category becomes implicated in the reasoning process, it is a *prototype* of the superordinate that is involved, rather than the superordinate category considered as an equivalence set. Osherson et al.'s model proposes that once the premise or premises have activated the superordinate, the resulting argument strength applies *equally* to all category members. The reasoning according to their model might thus be explicated as involving the following two stages:

1. All vultures have sesamoid bones; therefore, all birds have sesamoid bones.

2. All birds have sesamoid bones, all ostriches are birds; therefore, all ostriches have sesamoid bones.

The first step in the reasoning will have variable strength, depending on premise typicality, but the second step is taken in their model to be always perfectly strong on logical grounds. (Step 2 has the classical form of a syllogism.) There is evidence however from Sloman (1998) that people do not always respect the logic of Step 2 (see also Hampton, 1982, for failures in the transitivity of category superordination). As was described in the introduction, Sloman (1998) showed that people still preferred arguments from a superordinate to a typical subset over arguments from the same superordinate to an atypical subset.

According to the proposed modification to the model, the reasoning would now proceed as follows:

1. All vultures have sesamoid bones; therefore, prototypical birds have sesamoid bones.

2. Prototypical birds have sesamoid bones; therefore, all ostriches have sesamoid bones.

Unlike Osherson et al.'s (1990) model, the middle term of the category-based induction in our model is the prototype for the category, and not the category as a whole class. The category is being represented as an intensional concept, and not as an extensional set of exemplars. (See Tversky & Kahneman, 1982, and Hampton, 1987, for additional evidence that through representing concepts in terms of their properties, people are prone to ignore logical constraints on set membership.) By using intensional representation, the quantification of the middle argument "birds have sesamoid bones" is left vague, and so Step 2 cannot be given maximum strength on logical grounds.

Step 1 is of variable strength and will depend, as before, on premise typicality. Step 2, however, will also be of variable strength and will now depend on conclusion typicality. Moreover, these two steps rely on the same general principle as the direct route for assessing argument strength—namely, the similarity between two concepts.

Osherson et al.'s (1990) model applies to a much wider range of inductive problems than those considered here. One phenomenon that would not be explained by a simple activation account such as that offered here is the diversity effect. Arguments using two premise terms from the same category as the conclusion are considered less strong if the two premise terms are similar than if they are diverse. Activation of the prototype concept for the superordinate, according to our proposal, would therefore have to be greater in the case of diverse premises. Possible mechanisms by which to achieve this would be to base prototype activation on a sum of similarity to category exemplars (as in Osherson et al.'s model) or, alternatively, to base prototype activation on a measure of feature overlap between the disjunction of the features of the two premises and the prototype, as in Sloman's (1993) model.

Asymmetry

As it stands, the proposed modification to Osherson et al.'s (1990) two-route model would predict that the final strength of the argument, due to the indirect route, would be the product of the strengths of the two links. However, that would imply symmetry in the strength of arguments when the premise and the conclusion terms are reversed. Some alternative account, therefore, remains to be given of the premise–conclusion asymmetry effect, assuming it to be reliable. One can look for an answer in two places—in the direct route and in the indirect route—either or both of which may introduce asymmetry.

Asymmetry in the *direct* route would arise if the underlying similarity between the concepts is itself asymmetric. There is evidence that similarity is greater for a pair of items when the more typical or salient term is the target. For example, Tversky (1977) reported that the similarity (as rated by U.S. students) of Cuba to the United States is rated as greater than the similarity of the United States to Cuba. If premise–conclusion similarity were asymmetrical in a similar way (and computed as the similarity of the conclusion term to the premise term), that could account for the effect of reversing the premise and the conclusion within an argument. However, as was noted in the introduction, Aguilar and Medin (1999) reported a failure to replicate asymmetries in similarity ratings.

A second locus for asymmetry would be in the indirect route. Note that to be consistent with asymmetry, our result implies that there must be both a premise and a conclusion typicality effect but that the premise typicality effect must be greater than the conclusion typicality effect. One way to achieve this would be for the generalization gradient around a category prototype, such as *bird*, to be broader than that around a subclass concept, such as *robin* or *ostrich.* Having narrow generalization gradients, individual concepts will activate the general category name only if they are close to it. Typical premises, such as *robin*, will activate *bird* strongly, but atypical premises will not. The effect of premise typicality is, therefore, very pronounced. Once activated, the category name generalizes more broadly across all members of the category. Although it will still activate typical conclusions more strongly than atypical conclusions, the gradient of the typicality effect will be much shallower.

Conclusion

In this article, we set out first to demonstrate the existence of a phenomenon previously claimed not to exist—namely, that the typicality of a conclusion category affects the judgment of strength in category-based induction. The phenomenon was demonstrated with two different procedures across three biological categories, and an explanation in terms of familiarity was discounted. Finally, the implications of the result were discussed. First, they appear to directly contradict the account given by Sloman's (1993) feature-based model for typicality effects, since (unless typical items should turn out to have fewer features) the explanation he provides for the typicality asymmetry effect predicts that less typical conclusions will have stronger argument strength. Second, a modification of Osherson et al.'s (1990) model was proposed in which the strength of an argument is propagated via the superordinate category, but through activation of the category prototype, rather than through the set of category members.

REFERENCES

Aguilar, C. M., & Medin, D. L. (1999). Asymmetries of comparison. *Psychonomic Bulletin & Review*, **6**, 328-337.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

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

Gelman, S. A., & O´Reilly, A. W. (1988). Children's inductive inferences within superordinate categories: The role of language and category structure. *Child Development*, **59**, 876-887.

Hampton, J. A. (1982). A demonstration of intransitivity in natural categories. *Cognition*, **12**, 151-164.

Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory & Cognition*, **15**, 55-71.

Hampton, J. A., & Gardiner, M. M. (1983). Measures of internal category structure: A correlational analysis of normative data. *British Journal of Psychology*, **74**, 491-516.

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

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 411-422.

Lorch, R. F., Jr., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 149-157.

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

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

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

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

Sloman, S. A., & Wisniewski, E. (1992). Extending the domain of a feature-based model of property induction. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 355-359). Hillsdale NJ: Erlbaum.

Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**, 327-352.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84-98). Cambridge: Cambridge University Press.

(Manuscript received March 7, 2003;
revision accepted for publication July 4, 2003.)