Why Does Similarity Correlate With Inductive Strength?

Uri Hasson (uhasson@princeton.edu)
Psychology Department, Princeton University
Princeton, NJ 08540 USA

Geoffrey P. Goodwin (ggoodwin@princeton.edu)
Psychology Department, Princeton University
Princeton, NJ 08540 USA

Abstract

It has been repeatedly demonstrated that a robust predictor of the strength of an inductive argument is the similarity between the categories that are the focus of the induction. In this paper we evaluate why similarity is associated with the strength of such arguments. On one view, category similarity makes an argument strong because similarity is partially determined by features that are common to both categories, and the existence of these common features provides reason to think that the conclusion is justified. On another view, increased similarity may reflect few differences, so that there are not many reasons to think that the conclusion is unjustified. We evaluate this issue by examining how engagement in inductive reasoning affects the perceived similarity between categories. Our findings suggest that people attempt to find reasons to disbelieve the hypothesis suggested by an argument. They consider differences when evaluating inductions that posit an affirmative contingency, and consider similarities when evaluating inductions that posit a negative contingency. This is done independent of whether the induction is presented in argument form or in the form of a conditional statement, and independent of whether one is evaluating the truth or falsity of the conditional statement.

Introduction

The ability to reason by induction is one of the tools that make it possible to increase knowledge. And one of the stronger predictors for the strength of an inductive argument is the similarity between the categories that are the focus of the induction. People tend to find inductive arguments that involve highly similar categories (see Argument 1, below) to be stronger than arguments that involve categories that are less similar, e.g., Argument 2, (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990):

Argument 1:
Premise: Robins use serotonin as a neurotransmitter.
Conclusion: Sparrows use serotonin as a neurotransmitter.
Argument 2:
Premise: Robins use serotonin as a neurotransmitter.
Conclusion: Geese use serotonin as a neurotransmitter.

Admittedly, similarity is not the only determinant of inductive strength. The nature of the projected property (e.g., "use serotonin as neurotransmitter") and its relation to the category in question can override similarity. For example, Smith, Shafir, and Osherson (1993) demonstrated that the conclusion "German shepherds can bite through barbed wire" is supported more strongly by knowledge that Poodles can bite through barbed wire than by knowledge that Dobermans can. A number of other factors have also been shown to affect the strength of single-premise arguments such as those above (see Heit, 2000 for a review). Nonetheless, premise-conclusion similarity has been continuously demonstrated to be a strong determinant of the strength of an induction, especially when there is not much knowledge of the categories in question or the property mentioned in the argument.

The question we examine here is *why* premise-category similarity is such a strong predictor of inductive strength. One intuitive answer has been proposed by Mill (1874, in Heit, 1997); "because *a* resembles *b* in one or more properties, then it does so in certain other properties". For example, one would be justified in assuming that if Tunas thrive in sunlight then Goldfish will too. The argument being, "because tuna and goldfish are similar in some respects, it seems plausible that they will be similar in terms of a novel property, *thrives in sunlight*, as well" (Heit, 1997). Mill's suggestion clearly focuses on one explanatory factor: those properties of entities *a* and *b* that 'resemble' each other. Nowadays, such properties may be referred to as shared properties.

While modern models of similarity still ascribe an importance to shared properties, they also attribute great importance to those features of the categories that are not shared (Tversky, 1977; Gentner & Markman, 1997). Such conceptions motivate a more detailed exploration of the precise link between similarity and inductive strength. Inductions could be made (or evaluated) on the basis of common properties, distinct properties, or both. While Mill's account attributes the strength of an induction to the salience of shared properties, it is possible that the evaluation of argument strength is at least partially influenced by salient differences between the categories.

The effect of differences and similarities is particularly relevant to cases in which inductions are being evaluated. For example, when evaluating the argument "Tunas thrive in sunlight. Therefore, Goldfish do too", one can be aware of the fact that Tunas are big and Goldfish are small, that Goldfish are pet-fish, while Tuna are not, and so on. Once the differences between the categories are salient, their relevance to the property in question can be determined.

And, if the relevance is small, one could quite likely conclude that there is no reason to assume that the argument would be false. Additionally, people find it easier to list differences for more similar items than for less similar ones (Gentner & Markman, 1997), and so even relatively similar items may afford differences. The evaluation of arguments might also make similarities salient, e.g., the fact that that both Tuna and Goldfish are fish. The relevance of such similarities could then be assessed with respect to the property in question. It is unclear which sorts of information are used to evaluate inductive arguments such as Arguments 1 and 2 above. It could be that people try and find features common to both categories in order to see whether there is a good reason to believe that the property mentioned in the argument would also be shared. On the other hand, the evaluation of the argument might focus on differences between the categories to see whether there are good reasons to doubt that the property mentioned is shared. It is unclear which model best captures human reasoning, and this has been a topic of many debates between philosophical accounts of the justification of beliefs (see Harman, 1986).

A parallel issue is the extent to which such evaluations have long-term effects on the conceptual representation of the categories in question. While it is evident that the strength of inductive arguments is strongly predicted by the similarity of the categories, it is currently unknown whether engagement in inductive reasoning can in turn affect the perceived similarity of the categories. Since induction can lead to the acquisition of new knowledge, it is quite likely that inductive processes may lead to changes in conceptual representation.

In this paper we evaluate which properties are made salient in the evaluation of categorical induction, and test factors likely to mediate this process. We evaluate whether people focus on common or distinct features in their evaluations, and whether this is mediated by the syntactic form of the induction (Study 1). We then examine whether evaluations of truth and falsity affect peoples' focus on common and distinct features (Study 2).

Study 1: Inductive Reasoning and Conceptual Change

This study examined whether participants reason from similarities or differences during the evaluation of simple inductions and whether the syntactic form of the conclusion prompted different kinds of considerations.

We evaluated the considerations utilized in the evaluation of inductions through a study consisting of two separate stages. In the first stage, participants evaluated how likely it was that certain statements were true. These statements were conditional statements such as *If motels have nonvariable insurance policies then hotels have nonvariable insurance policies*. Such statements were used because previous research has demonstrated a near-perfect correlation between the strength of inductive arguments and the likelihood of those arguments when transposed into the form of conditional statements (Hadjichristidis et al., 2001). The second stage of the study was conducted about twenty-five minutes after the first stage. In this stage, participants rated the similarity of the categories mentioned in the statements (e.g., *how similar are motels and hotels*). This was used to see whether evaluation of arguments causes a long-term change in the perception of the categories, rather than a transient one.

We compared the similarity ratings given by those participants who had evaluated the statements to similarity ratings provided by a control group (who had not evaluated the statements beforehand). If the evaluation of conditionals prompted a search for common features then similarity ratings given by participants who evaluated the conditionals would be equal to, or higher than ratings provided by the participants in the control group. In contrast, if the evaluation of inductions highlights differences between categories, then participants in the experimental group should rate the categories as less similar than participants in the control group.

Our second interest was in the role that argument form may play in the sampling of properties. As mentioned earlier, the similarity ratings given after the evaluation of the inductive arguments reflect the accessibility and weighting of common and distinct properties. We wanted to know whether the evaluation of inductive arguments simply *reflects* knowledge about the categories in question, or whether it is the case that inductive arguments actually *frame* a specific hypothesis for evaluation, which in turn prompts a selective sampling of properties. Different argument forms may differentially weight common and distinct properties. For example, the evaluation of an induction in the form of an *If P then Q* conditional may lead to consideration of properties common to both categories, and therefore increase the similarity of the categories mentioned in *P* and *Q*. Note that this does not mean that the evaluation of any induction would increase the similarity of the premise and conclusion categories. Quite the contrary: evaluating an induction of the form *If P then not-Q,* (e.g., *If motels have nonvariable insurance policies then hotels will not have nonvariable insurance policies)* could make distinct properties salient. In other words, it could be that when evaluating inductions, participants do not consider knowledge of the categories in a context independent way, but are biased towards confirming a hypothesis suggested by the statement they are evaluating. If different forms of argument result in different weighting of common and distinctive features, then the consideration of such arguments would be followed by different patterns of similarity judgments.

Method

Participants. Eighty-Eight Princeton University undergraduates participated in the study for course credit

Design. We constructed twenty-five statements that depicted possible contingencies between two categories; e.g, *If Cows have stenozoidal cells, then Horses will also have stenozoidal cells*. All properties were 'blank' properties for

which it was expected that participants would not have much knowledge. Such properties are often used to isolate the effects of similarity and related factors. One third of the participants read conditional statements in which the antecedent and the consequent were affirmative ('Affirmative conditionals' henceforth). Another third read statements in which the antecedent was affirmative and the consequent was negated ('Negative conditionals' henceforth); e.g., *If Cows have stenozoidal cells, then Horses will not have stenozoidal cells*. Finally, one third of the participants did not read any statements at all. In the second stage of the study participants rated the similarity of the relevant categories (e.g., Cows and Horses).

Procedure. Participants in the two experimental groups were presented with booklets containing twenty-five statements. For each group the statements were arranged in two random orders. Participants were asked to rate "how likely it is that a given statement is true". They were not told that another section would follow. Ratings were made on a scale of 1 (*not at all likely)* to 10 (*very likely*). The second stage was administered twenty-five minutes later. In this stage, participants in all three groups received booklets for making the similarity judgments. Each pair of categories was printed separately on a line, and participants were asked to rate the similarity of the terms, e.g., "How similar are Cows and Horses?" The scale ranged from 1 (*not at all similar*) to 10 (*very similar*). The order of the terms in the question was identical to their order of appearance in the conditional statement.

Results and Discussion

Two participants in the control group were removed because there was no variance in their responses, and the task was not one in which this type of response is reasonable. Table 1 presents the mean ratings for the likelihood of the statements (out of a possible 10), and the mean similarity ratings given by the three participant groups.

Table 1. Mean likelihood ratings for experimental groups and subsequent similarity judgments.

Statement Evaluated	Likelihood	Similarity
Affirmative Conditional	5.07	5.36
Negative Conditional	4.27	5.93
None (Control)		5.77

The mean likelihood ratings were near the middle of the scale for both experimental groups, indicating that on the whole, there were no strong grounds for either accepting or rejecting the statements. To evaluate whether participants were seriously considering the statements, we examined the correlation between the likelihood of the statements and the similarity of the categories as rated by the control group. The correlation (Pearson's R) between the likelihood of the affirmative conditionals and the similarity of the categories was .51 ($p < .01$). The correlation between the likelihood of the negative conditionals and premise-conclusion similarity was -.46 ($p < .05$). The correlation measures indicate that participants were sensitive to the similarity of the premise and conclusion categories, and therefore had taken the task seriously.

Similarity ratings given after the evaluation of conditionals with negated consequents were higher than those given following the evaluation of conditionals with affirmative consequents; $t_s(57) = 2.3$, $p < .05$, $t_i(24) = 4.2$, $p < .001$. This finding addresses two of the motivations for the study. First, the evaluation of negatives and affirmatives seemed to have prompted different considerations. Second, the data are consistent with the possibility that the evaluation of conditionals with negated consequents focused participants on features common to both categories, whereas the evaluation of affirmative conditionals focused them on distinct features (differences).

There was no significant difference between the similarity ratings of the control group and those given by participants who evaluated the negative conditionals (ps > 0.5 by subjects and items, Bonferroni). This is consistent with the notion that the relative weighting of common and distinct properties did not differ between these groups. The difference between the ratings in the control group and ratings in the group evaluating the affirmative conditionals was significant by items, $t_i(24) = 2.5$, $p < .05$ (Bonferroni), but not by subjects ($p > 0.1$).

In sum, the study demonstrates that the evaluation of inductions in the form of affirmative and negative conditionals prompted different sorts of consideration of the categories in question. The evaluation of affirmatives resulted in lower similarity ratings than those given by the participants evaluating negatives, even though the similarity ratings were made twenty-five minutes later.

Given the rather non-intuitive nature of the results, we conducted a replication where the materials were presented in the form of standard inductive arguments (e.g., Cows have stenozoidal cells. Therefore horses have [do not have] stenozoidal cells). Participants (N = 20) rated the strength of the arguments and 25 minutes later judged the similarity of the categories. The mean similarity ratings given by participants who had evaluated arguments with affirmative and negative conclusions were 5.51 and 6.53 respectively. The difference between these two ratings was reliable; $t_s(18) = 3.1$, $p < .01$, $t_i(24) = 7.4$, $p < .001$. These results indicate that the evaluation of conditional statements and inductive arguments prompted similar sorts of considerations. Most important, in both cases it seems as if participants were recruiting information to assess whether the contingency implied by the argument or the statement is *incorrect.*

Valid arguments are those in which the conclusion must be true if the premises are true. Otherwise, an argument is invalid. Logically speaking, inductive arguments are invalid, but vary in their strength – i.e., in the support that the premises provide for the conclusion. Formally, arguments are judged in terms of a relation between the premises and the conclusion, not in terms of their relation to a possible state of affairs in the world. In contrast,

statements can be evaluated in terms of how accurately they capture, or may capture, a possible state of affairs in the world. For sentences with simple logical connectives, people are able to state which possibilities hold if the statement is true, and which hold if the statement is false. In the second study we evaluated whether focusing participants on these different possibilities could prompt different considerations of common and distinct features. We used the same conditional statements as in Study 1, and asked some participants to evaluate how likely it is that these statements were true, and asked other participants to evaluate how likely it is that they were false. The main purpose of the study was to see whether asking participants to evaluate truth and falsity would affect which properties would be sampled when evaluating the arguments.

We describe here possible outcomes for affirmative conditionals. The results of Study 1 are consistent with the notion that participants considered distinct features when asked to evaluate the likelihood that an affirmative conditional is true. Asking participants to evaluate the likelihood that an affirmative conditional is false might lead to a different sampling of properties. Participants may attempt to counter the claim that an affirmative contingency is false by searching for common properties. In this case, we would expect subsequent similarity ratings to be relatively higher than those given when considering the truth of statements. Another possibility is that participants will evaluate the falsity of conditional statements using the same sort of evidence recruited to evaluate their truth (i.e., differences). In this case, we would expect subsequent similarity ratings to be quite similar to those given when considering the truth of statements.

Study 2: Evaluations of Truth and Falsity

Study 2 examined these possibilities. Specifically, we investigated the considerations used to evaluate the truth of inductions and those used to evaluate the falsity of inductions. This study also aimed to replicate the findings of Study 1.

One difference between this study and the previous one was that each participant evaluated two blocks of statements: a block of affirmative conditionals, and a block of negative conditionals (order was counterbalanced). The statements differed only in the valence of the conditional's consequent. If negative conditionals prompt the generation of similarities and affirmative conditionals prompt the generation of differences, then certain transfer effects between the blocks are predicted, though the patterns may depend on whether truth or falsity is evaluated. The situation is quite clear for evaluations of truth:

1. Since the evaluation of negative conditionals makes *similarities* salient, affirmative conditionals should be rated as more likely to be true when they are evaluated after negative conditionals than when evaluated before them.
2. Conversely, since the evaluation of affirmative conditionals makes *differences* salient, negative conditionals should be rated as more likely to be true when they are evaluated after affirmative conditionals than when evaluated before them.

If the evaluation of falsity is based on the same considerations as the evaluation of truth, then evaluating affirmative conditionals should highlight similarities and evaluation of negatives will highlight differences. However, we did not make specific predictions as it was unclear which considerations were relevant to evaluations of falsity.

Method

Participants. Ninety-two Princeton University undergraduates participated in the study for cash payment.

Design, Materials and Procedure. We used the twenty-five materials employed in the first study. Participants completed one block in which they evaluated the likelihood of affirmative conditionals and one block in which they evaluated the likelihood of the corresponding set of negative conditionals. The order of blocks was counterbalanced between participants. In addition, half of the participants evaluated how likely it was that the statements were true and half evaluated how likely it was that the statements were false. The mixed design then was a 2 (Order) X 2 (Evaluation: true vs. false) X 2 (Valence: affirmative consequent vs. negative consequent) with Order and Evaluation manipulated between participants. Between the two evaluation blocks participants completed a filler task, which took about 15 minutes to complete. Following this stage, participants completed another filler task, and then rated the similarity of the categories that appeared in the statements. The procedure was identical to that detailed for Study 1.

Results and Discussion

Evaluations of Likelihood. The mean ratings for the likelihood of the statements in the different conditions are presented in Table 2. An analysis revealed strong transfer effects between blocks, so we present the data for each block separately. Note that the conditional statements presented in the first and second block had different valences. For example, if an affirmative conditional was evaluated in the first block then a negative conditional was evaluated in the second block. The Difference measure captures the carryover effects from the first block.

Table 2. Mean likelihood ratings as a function of task, valence and block position.

Conditional Evaluated	Evaluation Task	
	Likelihood of Being False	Likelihood of Being True
Affirmative in block 1	5.8	4.7
Affirmative in block 2	5.1	6.2
Difference	*0.7*	*-1.5*
Negative in block 1	6.4	3.9
Negative in block 2	5.7	3.9
Difference	*0.7*	*0*

Due to the transfer effects, we present the analysis of each block separately. The analysis of mean likelihood ratings for the first block revealed a main effect of the Evaluation task, as statements were rated as more likely to be false than to be true (M = 6.09 vs. 4.33), $F(1,90) = 72$, $p < .001$. If participants evaluated falsity by evaluating the truth of a conditional statement with an opposite consequent-valence, then no effect of Evaluation would have been found. In that case, ratings should have mirrored each other. For instance, for ratings in the first block, the likelihood of the affirmative form in the 'false' condition (5.8) should have been similar to the ratings for negative conditionals in the 'true' condition (3.9). Similarly, the likelihood for the negative form in the 'false' condition (6.4) should have been similar to the ratings for affirmative conditionals in the 'true' condition (4.7). But, as demonstrated in Table 2, this was not the case. This indicates that participants did not evaluate the falsity of such conditionals by assessing the likelihood that a conditional with the opposite consequent-valence is true. However, some inverse relations should hold between the likelihood of truth and falsity, and the expected inverse relation between the likelihood of a statement being true and its likelihood of being false produced the expected Evaluation X Valence interaction, $F(1,90) = 12.2$, $p = .001$.

The analysis of the mean likelihood ratings for statements presented in the second block revealed that affirmative conditionals were rated as more likely than negative conditionals, $F(1,88) = 12.98$, $p = .001$. As with the first block, we found the expected Evaluation X Valence interaction $F(1,88) = 40.57$, $p < .001$.

Observing the pattern of transfer effects enables us to evaluate our predictions. For evaluations of truth, affirmative conditionals were rated as more likely to be true when they were evaluated after negative conditionals than when evaluated before them (M = 6.2 vs. 4.7, $p < .05$, Bonferroni). However, we did not find the expected increase for the likelihood of negative conditionals following the evaluation of affirmative conditionals -- negative conditionals were rated as equally likely in both blocks. For evaluations of falsity, affirmative conditionals were rated as *less* likely to be false when they were evaluated after having evaluated negative conditionals in the first block (M = 5.1 vs. 5.8, $p < .05$, Bonferroni). This indicates that evaluating the falsity of negative conditionals highlighted similarities between the categories. Negative conditionals were rated as *less* likely to be false when they were evaluated after affirmative conditionals (M = 5.7 vs. 6.4, $p < .05$, Bonferroni). This indicates that evaluating the falsity of affirmative conditionals highlighted differences between the categories; once differences are made salient, it becomes less likely that a negative contingency is false.

The transfer effects between the blocks strongly support the notion that evaluation of affirmative conditionals is biased towards the consideration of distinct properties and that the evaluation of negative conditionals is biased towards the evaluation of common properties. The evaluation of truth or falsity does not determine whether similarities or differences will be searched for. The crucial observations are found in the similarity ratings.

Evaluations of Similarity. Note that all participants in this study evaluated both negative and affirmative conditionals. But, as demonstrated, considerations employed in the first block continued to affect later evaluations. The mean similarity ratings are shown in Table 3.

Table 3. Mean similarity ratings as a function of the evaluation task and valence of conditionals in the first block.

	Evaluation Task	
Conditional Valence	Falsity	Truth
Negative	5.75	6.07
Affirmative	5.50	5.22

We conducted a 2 (Valence) X 2 (Task) between-subjects ANOVA. The ANOVA revealed a main effect of Valence. Similarity ratings were higher after the evaluation of negatives (M = 5.89) than after the evaluation of affirmatives (M = 5.36), $F(1,84) = 5.50$, $p < .05$. This effect was also highly significant in the item analysis, which also revealed a significant Valence X Evaluation interaction, $F(1,24) = 12.4$, $p < .01$. The strong effect of valence provides additional support for the notion that the evaluation of affirmative and negative conditionals prompts different consideration of commonalities and differences. The nature of the task, be it evaluation of falsity or truth perhaps modifies this tendency, but does not eliminate it completely.

General Discussion

We set out from the general finding that the strength of inductive arguments is strongly predicted by the similarity of the categories. However, similarity can vary with common and distinct features, and it was unclear which types of properties are actually sampled when evaluating inductive arguments.

We examined which types of properties were accessed during the evaluation of induction by asking participants to evaluate the strength of arguments and then rate the similarity of the categories mentioned in the arguments. Our findings indicate that people are highly strategic in their evaluation of even simple inductive arguments. People seem to search for features that would result in doubting the hypothesis put forward by the argument. For example, they seem to sample distinct features when evaluating arguments such as *Cows have stenozoidal cells. Therefore, horses have stenozoidal cells.*

One can ask whether this is a reasonable strategy to use when establishing credibility. This issue has been taken up by Gilbert Harman, (1986), who contrasts two possible principles for justifying belief: on one principle, people only believe those things for which they have proper justification on the basis of other beliefs. If no justifications exist then there is no reason to believe. On another view, beliefs are

kept not due to justification by other beliefs, but due to consistency with other beliefs. On this view, beliefs will only be rejected if there are special reasons to doubt them (lack of justification not being such a reason).

Our results are more consistent with the latter of these two notions. Rather than looking for reasons to think that a certain hypothesis is justified by supporting facts, people appear to search for special reasons to reject the hypothesis put forward.

It would be interesting to know whether the tendency to notice differences when evaluating inductions is an ability that develops early in life. A large body of research has demonstrated that children are more likely to project a property from one entity to another when the two entities are similar. For example, Carey (1985) demonstrated that the strength of inductions increased as premise-conclusion similarity increased, and sensitivity to this relationship increased with age. In light of the present findings, these similarity effects (or 'similarity correlates') in children's induction appear not to have been fully decomposed. As our discussion to this point demonstrates, children might be aware of the salience of common features, distinct features or both. And it could be that there are qualitative differences between children and adult reasoning in this respect. One plausible hypothesis is that children's induction is initially focused on commonalities, particularly perceptual ones, but becomes more able to accommodate differences as cognitive capacity increases.

Finally, the results also highlight long-term effects of reasoning. The evaluations that are generated during inductive reasoning cause conceptual change in the similarity of the relevant categories, and this change is relatively long term. Given that induction is an efficient tool for knowledge acquisition, such changes are not surprising.

Acknowledgements

This research was supported in part by a grant from the National Science Foundation (BCS-0076287) to study strategies in reasoning. We thank Philip N. Johnson-Laird for his advice.

References

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

Gentner, D., & Markman, A. B. (1997). Similarity mapping in analogy and similarity. *American Psychologist, 52*(1), 45-56.

Hadjichristidis, C., Stevenson, R. J., Over, D. E., Sloman, S. A., Evans, J. S. B. T., & Feeney, A. (2001). On the evaluation of If p then q conditionals. *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*. (pp. 381-386). LEA, Hillsdale, NJ.

Harman, G. (1986). *Change in view: Principles of reasoning*. London, England: Bradfod/MIT Press.

Heit, E. (1997). *Features of similarity and category-based induction.* Paper presented at the Proceedings of the Interdisciplinary Workshop on Categorization and Similarity, University of Edinburgh.

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

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

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

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition, 49*, 67-96.

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