Can similarity-based models of induction handle negative evidence?

Daniel Heussen (Daniel.Heussen@psy.kuleuven.be)

Wouter Voorspoels (Wouter.Voorspoels@psy.kuleuven.be)

Gert Storms (Gert.Storms@psy.kuleuven.be)
Department of Psychology, Tiensestraat 102, 3000 Leuven, Belgium

Abstract

Even if we don't like it, we often face counterexamples to the inferences we have made or would like to make. With the exception of the SimProb model (Blok, Medin & Osherson, 2007), models of inductions to date have predominantly focused on the relevance of positive evidence to the inference process. Here we provide data from single and double premise arguments in a category-based property induction task using positive and negative evidence. A simple similarity model, the Similarity-Coverage model (Osherson et al., 1990) and the SimProb model are tested on negative and mixed evidence arguments.

Keywords: Induction; Negative evidence; Similarity

The relevance of negative evidence

Ever since Hume, induction has been an area of immense research efforts in philosophy (e.g., Goodman, 1955; Hempel, 1966; Lipton, 2004), psychology (e.g., Blok, Medin, & Osherson, 2007; Heit, 2000; Osherson et al., 1990; Rehder, 2009; Rips, 1975; Sloman, 1993) and cognitive science (e.g., Kemp & Tenenbaum, 2009) in general. Among the prominent questions studied have been: What is the logical basis for induction? What role does prior (semantic) knowledge play in inductive reasoning? Why are some kinds of fact more easily projectable than others? And how should we model inductive inference? Despite these extensive efforts less is known on the influence of negative evidence in induction.

Negative evidence, however, is ubiquitous in everyday reasoning. In some circumstance, evidence may go against our established views. Your favorite restaurant serves you a bad meal, your friend, that is always late, shows up on time and your oh so reliable car won't start. In other instances, you might be making a new inference with both positive and negative evidence present. You check out a new restaurant and receive a great starter and desert but a burned steak and overcooked vegetables. Negative evidence in category-based property induction is defined here as evidence from an instance of the conclusion category that does not possess the to-be-projected property. In other words the evidence constitutes a clear counterexample of something possessing the to-be-projected property. The questions we would like to address here are: How does negative evidence affect our generalizations? What determines the relevance of negative evidence? How do we combine evidence to reach a conclusion?

In research on induction involving positive evidence, Rips (1975) found that the similarity of the evidence to the conclusion influences its relevance. People are more willing to generalize the attribution of a property from a robin to a sparrow than from an eagle to a sparrow because robins and sparrows are more similar. Models of induction involving positive evidence have tried to capture this intuition. The similarity coverage model for instance uses the maximum similarity between premises and conclusion as one component to their model (Osherson et al., 1990). Similarly Sloman's (1993) feature model uses the overall match in the number of features between the premises and the conclusion as a determinant of argument strength. The SimProb model (Blok, Medin, & Osherson, 2007) turns similarity between premises and conclusion into probabilities and uses those to determine argument strength.

The question we are addressing here is where similarity also determines the relevance of negative evidence. If similarity functions in the same way for positive and negative evidence in determining whether a piece of evidence is considered to be relevant to the conclusion, then existing models of induction based on similarity should be able to handle arguments involving negative evidence. To our knowledge, the SimProb model (Blok, Medin, & Osherson, 2007) is the only model explicitly designed to handle negative evidence. Other models require some adaptation to handle the intuition that the belief in a proposition should decrease with the encounter of negative evidence.

A second question of importance when modeling induction is how to combine the evidence. One approach might be to simply add to argument strength for positive evidence and subtract for negative evidence. Alternatively as the SimCov (Osherson et al., 1990) and the SimProb model (Blok, Medin, & Osherson, 2007) suggest, one could assign the greatest importance to one premise by virtue of its similarity to the conclusion for instance and adjust the resulting argument strength in accordance with the remaining evidence. Furthermore the manner in which the second premise exerts its influence can be implemented in different ways. The SimProb model suggests a weighting by similarity to the first premise. The SimCov model uses the relative positions of the premise categories in a conceptual similarity space to determine the influence of additional premises. These are only a few examples of the various possibilities to combine data, but they highlight the complexity of the issue.

The aim here is to test whether similarity based models of induction are able to handle negative evidence in a category-based property induction task. We present data from an induction task involving single and double premise arguments with positive and negative evidence and fit three models. In the next section we'll describe in more detail the three models used.

Similarity-based models of induction

We evaluated three models, each relying essentially on similarity to predict the strength of an argument. The models differ in how information is combined in arguments with two or more premises and in the implementation of negative evidence premises. The first model is a simple similarity based model (Sim). The second model is the similarity-coverage model (SimCov) as proposed by Osherson et al. (1990). In the present study, we adapted the model to account for negative evidence. The third model is the similarity-probability model (SimProb; Blok, Medin, & Osherson, 2007).

The Sim model

In this model the strength of the argument is directly related to the similarity of the conclusion category and the premise category (or categories). Formally, the argument strength S_c of an argument with conclusion c and a set of premises then is:

$$S_c = \sum_{p=1}^{n} e_p \, sim_{cp}$$

where sim_{cp} is the similarity between the conclusion category and the category of premise p and e_p indicates whether the premise is positive or negative (respectively $e_p=1$ or $e_p=-1$). Note that in this expression similarities are combined in a very straightforward manner, summing them (or subtracting, depending on whether it's a positive or a negative premise) across the number of premises.

The SimCov model

In the SimCov model, the strength of an argument depends on two components. A similarity component captures the similarity between premise and conclusion categories, and thus the relevance of the premise. The coverage component captures the idea of how much of the nearest superordinate category containing both premise and conclusion categories is covered by the premise(s). We modified the model to account for negative evidence by making the similarity of a premise and a conclusion category negative when the premise is negative.

Formally, the argument strength according to the SimCov model is a weighted sum of the similarity and the coverage component:

$$S_c = \alpha \times similarity_c + (1 - \alpha) \times coverage_p$$

where α is a free parameter determining the relative weight of each component. The similarity component represents the similarity between premise and conclusion category. In case of multiple premises, the similarity component is equal to the premise category that is most similar to the conclusion category. As in the previous model, when the most similar premise category is in a negative premise, the similarity is negative.

The coverage component is calculated as follows:

$$coverage_p = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \max\left(\mathrm{sim}_{\mathrm{p_1 i}}, \mathrm{sim}_{\mathrm{p_2 i}}, \ldots, \mathrm{sim}_{\mathrm{p_n i}}\right)$$

where i is an element of a relevant comparison set and N is the size of that set. The comparison set consists of known members of the nearest superordinate category containing both premise and conclusion categories. The coverage term implements the diversity principle (Carey, 1985). In a double positive premise argument, the more diverse the two premise categories are, the larger the coverage term will be – the more the nearest superordinate category is "covered" by the premise categories. Again, when the most similar premise category is in a negative premise, the similarity is negative in the expression.

The SimProb model

In the simprob model, inductive reasoning is considered as a conditional probability judgment. Given a certain prior belief about something, the evidence considered will update this prior belief. Formally, the belief update elicited by the premise a is given by:

$$P(c|a) = P(c)^{\alpha}$$

with

$$\alpha = \left(\frac{1 - sim_{ca}}{1 + sim_{ca}}\right)^{1-P(a)}$$

When there are two premises, the most relevant premise a (the premise that would influence the prior belief the most) is combined with the lesser relevant premise in the following way:

$$P(c|a,b) = P(c|a) + \begin{bmatrix} \left(1 - P(c|a)\right) \times \\ \left(1 - sim_{ab}\right) \times \\ \left(P(c|b) - P(c)\right) \end{bmatrix}$$

There are elegant symmetrical expressions to implement negative evidence (see Blok et al., 2007, for details). The basic idea is that the probability of a negative premise is 1 minus the probability of the same but positive premise, and that similarity between two premises will raise the posterior probability of the conclusion instead of decreasing it.

The SimProb model makes use of prior beliefs regarding the premises and conclusion. In the present study, we use blank properties. Following Blok et al., (2007) in their handling of blank properties, we use a uniform and low prior probability (fixed at .2) for all premises and conclusions.

An obvious parallel between the three models is that they all rely heavily on similarity to account for argument strength. There are differences however, in how similarity is used and – for arguments with multiple premises – how premise information is combined. The Sim model simply adds and subtracts similarities in the multiple premise case. SimCov picks the most relevant premise based on similarity and discards the similarity of the other premise. SimProb picks the most relevant premise, updates the conclusion probability and then modifies the resulting probability according to the less dominant premises.

Present research

The primary goal of this study was to see whether models that use similarity as a determinant of relevance of the evidence are able to handle negative evidence. To that end, we first established what influence negative evidence has on argument strength. We then tested a simple similarity model (the Sim Model), that only takes similarity into account, the SimCov model (Osherson et al., 1990) that also considers the coverage of the conclusion category and the SimProb model (Blok, Medin, & Osherson, 2007), that was specifically designed to be able to handle negative evidence.

The models are evaluated on data from a standard category-based property induction task using properties that participants are likely to have very little knowledge about. The properties are projected from either one or two exemplars to another exemplar of the same category. Participants are asked to judge how likely the conclusion is given the premises, for instance, how likely is it that magpies have a syrinx given that parakeets have a syrinx? The models are tested on four kinds of arguments:

Single Positive:
Parakeets have a syrinx.
Magpies have a syrinx.

Single Negative:
Parakeets **do not** have a syrinx.
Magpies have a syrinx.

Double Positive:
Parakeets have a syrinx.
Penguins have a syrinx.
Magpies have a syrinx.

Mixed Positive & Negative:
Parakeets have a syrinx.
Penguins **do not** have a syrinx.
Magpies have a syrinx.

Note that in the mixed arguments, the negative premise was always the premise presented second.

Method

Participants 76 students from the University of Leuven, Belgium, participated in the study. Participants received course credits in return for participation.

Design Two groups of participants rated the inductive strength of 40 target and 14 filler arguments. Filler items were arguments that were clearly true or false. One group evaluated 20 single positive arguments and 20 mixed positive and negative premise arguments. Fillers for this group consisted of single and double positive arguments. The other group evaluated 20 single negative premise and 20 double positive premises arguments with fillers being single positive and mixed positive and negative premises. The exemplars and properties used were identical for the two groups matching the characteristics across positive and negative arguments.

Materials To create arguments, we selected exemplars from four animal categories (i.e., birds, fish, insects & mammals) from the Leuven Concept Norms (DeDeyne, et al., 2008). For each category, the norms contain exemplars generated by participants as well as pair-wise similarity ratings between them. The norms also contain typicality ratings for each exemplar. Exemplars of the two premises and the conclusion were matched for typicality across the single and double premise arguments. The to-be-projected properties were biologically plausible blank properties. For each animal category we selected five kinds of characteristics (i.e., anatomical, behavioral, developmental, metabolic, necessity) that people were likely to have little knowledge about (e.g., Robins require amylase for their digestion). The task was administered in form of a questionnaire. The first page contained a description of the task with the instruction and an example argument. This was followed by 54 arguments starting with 3 warm-up fillers. The remaining 11 fillers were evenly distributed across the items. One random order of items and its reverse was used.

Procedure The induction task was presented as part of a battery of test. Students participated in a large group and took no longer than 10 minutes to complete the task.

Results

Preliminary Analysis Five participants were excluded from the analysis due to a lack of variance in their responses. In a subsequent reliability analysis, the two groups showed high consistency in their responding (Cronbach's alpha of .88 and .95). The data were averaged across participants and subsequent analyses were carried out on the items.

Manipulation Check Each of the 40 target items appeared once with positive and once with negative evidence. Of these, 20 items were single premise and 20 were double premise arguments. Figure 1 shows the average argument strength across those four conditions.

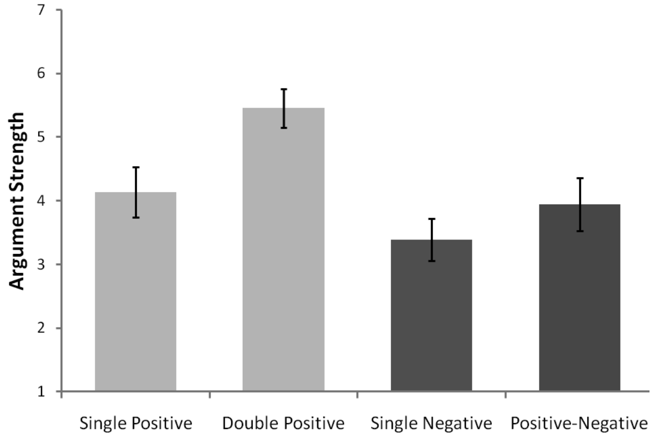

Figure 1: Argument strength for all four types of argument. Error bars are 95% CI.

Arguments containing negative evidence (darker bars) were rated lower in argument strength than those with positive evidence. For positive and negative evidence, arguments having two premises increased argument strength. Note though that in the mixed positive-negative premise arguments the increase in argument strength is due to the addition of a positive rather than negative premise.

The data were submitted to a 2 × 2 mixed factorial analysis of variance with type of evidence (contains negative evidence vs. does not contain negative evidence) as repeated measure and type of argument (single vs. double premise arguments) as between subjects factor. Although the data suggested that adding a positive premise has a greater effect if the first premise is positive as opposed to negative, the interaction between argument type and evidence type was not significant ($F(1, 38) = 3.2$, $p = .08$). Both main effects of type of evidence ($F(1, 38) = 27.8$, $p < .001$) and type of argument were significant ($F(1, 38) = 38.3$, $p = .001$). Single negative premise arguments were rated weaker than single positive premise arguments ($t(19) = 2.2$, $p < .05$). Similarly mixed positive-negative premise arguments were judged less strong than those with two positive premises ($t(19) = 5.9$, $p < .05$). Adding a positive premise to either a positive ($t(38) = 5.2$, $p < .05$) or a negative premise ($t(38) = 2.1$, $p = .05$) increased argument strength.

The data confirmed the intuition that negative evidence should have an adverse effect on argument strength. Arguments involving negative evidence were rated lower than those with positive evidence. For positive evidence, we also found a monotonicty effect (Nisbett, et al., 1983); more premises led to stronger arguments.

Modeling preliminaries In order to evaluate the model fits, we use the correlation between the averaged observed and predicted argument strength within each condition. To derive predicted values from the models, we extracted pair-wise similarity ratings between items from the Leuven Concept Norms (De Deyne, et al., 2008). Although the SimProb model provides predicted values in terms of conditional probabilities the other two models do not and we therefore do not make any claims about the scales of the predicted values and will not discuss differences between the models in those terms.

In terms of model parameters, the Sim model does not contain any parameters. The SimCov model uses the alpha parameter to determine the relative influence of its two components (i.e., the similarity component and the coverage component). Figure 2 presents model fits (i.e., correlations between predicted and observed) across the whole range of the alpha parameter. In all four conditions a reduction in the alpha parameter led to a reduction in fit indicating that the coverage term did not play a role. Consequently we fixed the alpha parameter at 1.

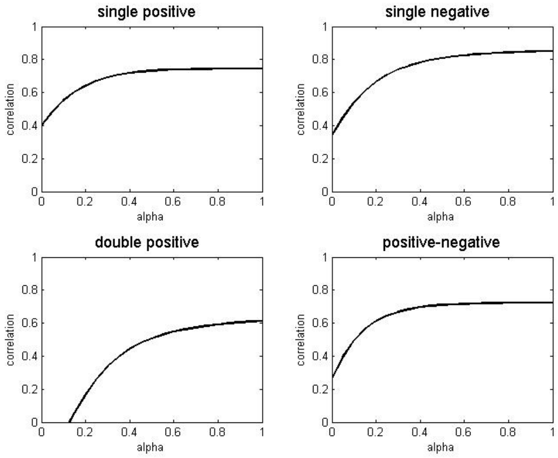

Figure 2: Model fits plotted against the whole range of the alpha parameter of the SimCov model in each condition.

The SimProb model requires prior probability judgments for the properties as input parameter to the model. Nevertheless, Blok et al. (2007) suggest that the SimProb model can handle arguments containing blank properties. They recommend using uniform and low prior probabilities, as this will ensure that the similarity component of their model will do most of the work. We therefore opted for uniform priors across premises and conclusion of .2.

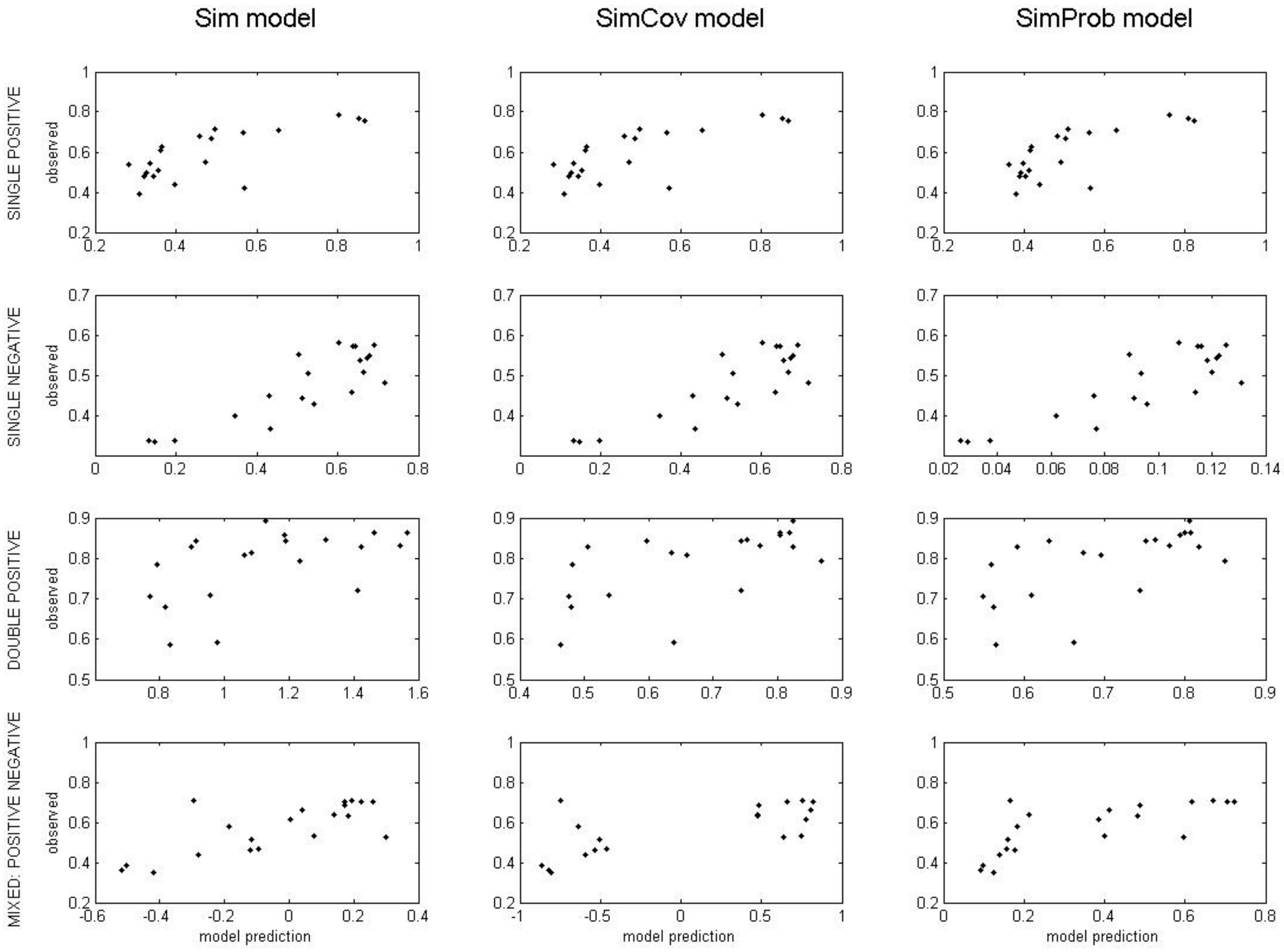

Figure 3: Scatter plots of observed against predicted for each model across single positive, single negative, double positive and mixed positive-negative argument.

Modeling results Figure 3 shows the scatter plots of the predicted versus observed values for each of the three models (columns) across the four types of argument (rows). All correlation coefficients were significant at $p < .05$ with n = 20. For single positive premises arguments (top row), the three models showed virtually identical results with a good fit of $r = .74$ for all three models. Looking at single premise arguments with negative evidence (2nd row), the models were equally capable at predicting participants' responses and even showed a better fit ($r = .85$). There was no difference in model predictions or fit across the three models. Thus for single premise arguments the three models can equally well account for argument strength involving positive and negative evidence.

The third row shows that for double positive premise arguments the three models differed in their predictions. The Sim model showed a somewhat weaker fit ($r = .53$) than the SimCov ($r = .61$) or the SimProb ($r = .62$) models. Applying a t-test to the Fisher's Z transformed correlation coefficients however showed that the difference was not significant ($t(17) = .56$, *n.s.*). Overall the fit of the models for double positive premise arguments was not as good as for single premise arguments.

Testing the fit for mixed positive and negative premise arguments (4th row) we find no difference between the models in terms of the correlation coefficient (Sim: $r = .75$; SimCov: $r = .73$; SimProb: $r = .73$). However the scatterplot shows that the SimCov model, unlike the other two, predicts two separate clouds of data points across the range of observed values. The human data clearly showed a continuous distribution across the whole range of possible values without two separate clouds. The difference in overall mean of each cloud in the predicted data seems to drive the correlation. This is due to the max function in the similarity component choosing the premise (positive or negative) that has the greater similarity and dropping the influence of the other premise. In contrast the Sim model and the SimProb model take both premises into account.

General Discussion

In making an inference, we have to determine whether a piece of information is relevant or not. For evidence in favor of our inference, theories of induction (Blok, Medin, & Osherson, 2007; Osherson, et al., 1990; Rips, 1975; Sloman, 1993) have suggested that the relevance is determined by the similarity between the evidence and the conclusion. In everyday reasoning, however, we often face at least some evidence that is not in line with our favored conclusion. Here we have tested whether models that use similarity to determine relevance are able to handle arguments involving negative evidence.

The model fits showed that for single premise arguments all three models were able to account for the data from both positive and negative premise arguments equally well. This indicates that the relevance of negative evidence can also be modeled using similarity. For double premise arguments all three models did a decent job with positive evidence. However, for mixed positive–negative premise arguments only the Sim and the SimProb model were able to account for the data. Although showing a good fit in terms of the correlation coefficient, the SimCov model showed a pattern of predicted values not reflected in human data. Taken together, two factors can account for the behavior of the SimCov model. First, with our data the coverage component of the SimCov model did not contribute to the prediction of argument strength. One reason for this might be that the generalizations in our arguments were to other exemplars rather than the category itself. Second, the similarity component only takes into consideration the most similar premise disregarding the other. If this happens to be the negative one, predicted values are low. Conversely if the max function selects the positive premise predicted values are high. Without an influence of the coverage terms two clusters of predicted values emerge.

The results from the double premise arguments again support the fact that similarity can be used to determine the relevance of negative just as well as positive evidence. However the results highlight that with several pieces of evidence it becomes important to consider how to model the combination of both positive and negative evidence. Differences in how the models combine the evidence make them better or worse candidates in modeling negative evidence with multiple premises. Disregarding one piece of evidence over another clearly does not resemble participants responses. However similarly a simple additive model like the Sim model becomes less realistic in the case of multiple premises of the same kind, evident in our double positive condition.

The aim of the present study was not to provide a new model of induction but to test whether similarity-based models of induction can handle arguments involving negative evidence. We have shown that similarity can indeed be used to model relevance of negative evidence. In addition, our data highlight the importance of taking all evidence into account. Models of induction that try to account for the influence of negative evidence will need have a specific mechanism to combine positive and negative evidence.

Acknowledgments

This research was supported by a postdoctoral research fellowship within the framework of international mobility awarded to the first author by KU Leuven under the supervision of Gert Storms.

References

Blok, S. V., Medin, D. L., & Osherson, D. (2007). Induction as conditional probability judgment. *Memory and Cognition, 35,* 1353–1364.

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

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavioral Research Methods, 40,* 1030-1048.

Goodman, N. (1955). *Fact, fiction, and forecast.* Harvard University Press.

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

Hempel, C. G. (1966). *Philosophy of Natural Science*. New Jersey, Prentice Hall.

Kalish, C. W. & Lawson, C. A. (2007). Negative evidence and inductive generalisation. *Thinking & Reasoning, 13*, 394-425.

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

Lipton, P. (2004). *Inference to the best explanation*. London, Routledge.

Nisbett, R. E., Krantz, D. H., Jepson, D., & Kunda, Z. (1983). The use of statistical heuristics in everyday reasoning. *Psychological Review, 90*, 339-363.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*(2), 185–200.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science, 33*, 301-343.

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

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