A mechanism to increase argument strength with 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, University of Leuven
Tiensestraat 102, 3000 Leuven, Belgium

Abstract

In most cases, if relevant, positive evidence raises and negative evidence lowers argument strength in induction. Previous research, however, has shown that it is possible to raise the argument strength of a single positive premise argument by introducing negative evidence (Heussen, et al., 2011). Here we test one possible mechanism for such an increase in argument strength. When people consider the positive premise they develop a set of hypotheses. Subsequently encountering a negative premise would render some of these hypotheses less likely and hence, if participants see the hypotheses as an exhaustive set, shift the probability to the remaining hypotheses to varying degrees. We test this idea by asking people to choose between four hierarchically nested conclusions across various conditions of evidence. The results are discussed in the light of models of induction.

Keywords: Induction; Negative evidence; Hypothesis Space

Negative evidence in induction

In its broadest sense, inductive reasoning can be defined as inference to an uncertain conclusion (Lipton, 2004). There are at least two main approaches to make such inferences. One is to generalize over time. "My car has always started, so it's reliable and will start today." There are a number of philosophical arguments why such an approach is not reliable (Russell, 1912), but owning an aging car will quickly teach you the same. Another approach is to generalize across instances of a category. "My car is a German car, so it's reliable and will start today." Although this approach similarly has its problems, it is certainly pervasive in everyday reasoning. As a consequence a large amount of research, particularly in psychology, has been looking at category-based property induction (see Feeney & Heit, 2007 and Heit, 2000 for overviews). In these studies, participants might be asked to judge how likely it is that *Bobcats* use serotonin as neurotransmitter given that both *Tigers* and *Cougars* do (Smith, Shafir & Osherson, 1993).

What is striking though is that in the majority of this research the focus has been on positive evidence, evidence that states that some entity has a particular property. And only very little is known about the influence of negative evidence, evidence that states that some entity does not have a particular property.

So what do we know about the influence of negative evidence on argument strength? In work on category-based property induction, negative evidence has been found to behave in a similar manner to positive evidence, however, with the reverse effect on argument strength (Heussen & Hampton, 2009; Heussen, Voorspoels, & Storms, 2010). For instance, just as for positive evidence the similarity between the negative evidence and the conclusion is a key predictor of the influence of the evidence on the conclusion. Similarly, in generalizations to the category, negative evidence instantiated by a less typical exemplar has a less detrimental impact on argument strength than from a more typical exemplar. Other studies have shown that in mixed premise arguments, containing both positive and negative evidence, the similarity between the contradicting evidence plays an important role with greater similarity between positive and negative evidence resulting in a stronger impact of the negative evidence (Blok, Medin & Osherson, 2007).

In line with intuition, these studies show that positive evidence raises argument strength—"Lions have enzyme x, tigers have enzyme x, how likely is it that cheetahs have enzyme x?"—and negative evidence lowers it—"Lions have enzyme x, tigers DO NOT have enzyme x, how likely is it that cheetahs have enzyme x?" Argument strength never goes against the "sign" of the evidence. Elsewhere we call this the Monotonicity Principle (Heussen et al., 2011).

Against the monotonicity principle

Contrary to the Monotonicity Principle, however, it has been found that in some circumstances it is possible to raise argument strength by introducing negative evidence (Heussen et al., 2011). Participants were asked to consider scenarios like the following: "Scientists have established that in certain brain regions the music of Dmitri Shostakovich elicits particular brain waves called alpha waves. Given that Shostakovich's music causes these alpha waves, you might wonder whether this similarly applies to other kinds of music like, for instance, the music of Bach." Participants were then asked to judge how likely it is that Bach's music also causes alpha waves? Participants then received additional information, that the same group of scientists found that the music of the hard rock band AC/DC does NOT elicit alpha waves in the brain. And again they were asked to make a judgment about whether Bach's music elicits alpha waves.

The results revealed a significant increase in argument strength from the first to the second judgment. Presenting people with exemplars from contrasting subcategories within the immediate superordinate category of all exemplars in the argument resulted in an increased endorsement of the conclusion. Counter the monotonicity principle, adding some kinds of negative evidence to a single positive premise argument increased rather than decreased argument strength.

In contrast, when presented with negative evidence from the same subcategory (e.g., classical music), as in "Haydn does not elicit alpha waves", participants showed a dramatic drop in argument strength. In line with the Monotonicity Principle negative evidence from the same subcategory resulted in a strong decline in argument strength.

The results of our study suggest that negative evidence can be used to highlight a relevant dimension or criterion on which to make the inductive leap. In the scenario above, general knowledge provides a range of commonalities and differences between Shostakovich's and Bach's music and the negative premise reduces these down to the relevant ones by highlighting respects of similarity (Medin, Goldstone & Gentner, 1993). In addition to influencing the inductive process at a similarity level, negative evidence may also impact the inductive process at the hypothesis level, affecting the kinds of hypothesis that are entertained. The search for commonalities and differences between the positive evidence and the conclusion exemplars leads to a range of hypotheses about why they may or may not share the particular property in question. By explicitly contradicting some of the generated hypotheses (e.g., not all music elicits alpha waves), negative evidence clearly helps in reducing the number of hypotheses. But not only that, evidence from concept learning suggests that negative evidence even constrains the generation of hypotheses already at the outset of learning (Houtz, Moore, & Davis, 1973). In sum, previous work suggests that negative evidence may in some circumstances constrain the process of induction both in terms of highlighting the relevant dimensions of similarity as well as constraining the generation and selection of hypotheses (Heussen et al., 2011).

Models of induction

Most models of induction, to date, have focused on the influence of positive evidence (e.g., Osherson et al., 1990; Rips, 1975; Sloman, 1993). To our knowledge, the SimProb model (Blok, Medin & Osherson, 2007) and Bayesian models of induction (e.g., Heit, 1998; Kemp & Tenenbaum, 2009) are the only two approaches to modeling induction that have explicitly incorporated the influence of negative evidence. All approaches, however, endorse the Monotonicity Principle—positive evidence raises and negative evidence lowers the credence in the generalization. The finding that negative evidence can in fact raise argument strength hence poses a serious challenge to all models of induction (Heussen et al., 2011).

The challenge to models of induction is twofold. First, the way models of induction currently incorporate negative evidence is limited to subtracting from existing argument strength in proportion to the influence of the evidence. Hence the effect of negative evidence ranges from no influence to a strong negative influence. This however does not allow for an increase in argument strength. Thus models require a mechanism to allow an increase in argument strength such that the effect of negative evidence can range from a strong negative to a slight positive influence. Second, models of induction need an a priori way to distinguish between negative evidence that reduces and negative evidence that increases argument strength. The former challenge is certainly easier to meet than the latter. Here we would, therefore, like to address the easier of these two questions and test a possible mechanism by which argument strength might increase when encountering negative evidence.

Rationale of the present study

The aim of the present study is to look at a possible explanation for the increase in argument strength in the above mentioned study. One way to introduce a possible increase in argument strength through negative evidence is to posit that people generate a finite set of hypotheses when encountering the positive premise. For instance, in the Shostakovich – Bach example above, people may develop a set of hypotheses for the 'elicitation of alpha waves' consisting of 'sound in general elicits alpha waves', 'only music does', 'only classical music does' or 'it only applies to Shostakovich'. Assuming a probability distribution over these hypotheses, would imply that, when excluding one hypothesis, there should be an increase in probability of those hypotheses further down in the hierarchy. In the example above, introducing negative evidence that excludes or reduces the likelihood of the 'music in general' hypothesis and as a consequence also the 'sound in general' hypothesis would hence lead to an increase in likelihood for the remaining two hypotheses.

In order to test this idea, we presented participants with either single positive or mixed positive and negative premise arguments. The arguments had four hierarchically nested conclusions ranging from the exemplar (e.g., Shostakovich) presented in the positive premise via its subcategory (e.g., classical music) and basic category (e.g., music) to its superordinate (e.g., sound). Using hierarchically nested hypotheses allowed us to enforce an exhaustive set of hypotheses. Participants were asked to choose the most general conclusion that is still acceptable based on the given information. Among the mixed premise arguments the negative evidence was varied by choosing exemplars at different levels of hierarchy. The negative premise either came from the same (e.g., Haydn) or a contrasting subcategory (e.g., AC/DC) as the positive evidence or it came from a different superordinate category (e.g., the sound of a falling screwdriver).

The idea of varying the exemplars in the negative premise was to elicit the shift in probability from the general to the most specific conclusion. More precisely, the single positive premise arguments constitute the baseline with the distribution across the conclusion predicted to show a monotonic decline in probability from exemplar to the superordinate level. Negative evidence from the same subcategory as the positive evidence should shift most of the probability to the exemplar level leaving the remaining three with little to no probability density. In contrast, negative evidence from a contrasting subcategory should result in a large probability for the subcategory level as conclusion as well as the exemplar level. Both basic level and superordinate level conclusions should receive little density. Furthermore, negative evidence from a different superordinate category should spread the probability making the basic level category more likely as a conclusion. This redistribution of the probability density should hence trace the level at which the negative evidence is pitched precisely one level lower in the hierarchy.

In order to show that this mechanism can lead to a significant increase in argument strength we need to focus on the specific conclusion at the subcategory level. In order for the mechanism to be sufficient to raise argument strength, the results need to show a significant increase in probability from the single to the mixed premise argument for the subcategory level conclusion, when the negative evidence is instantiated by a contrasting subcategory exemplar. In addition the usual detrimental influence of negative evidence should be observed in a significant decrease in probability for that conclusion from the single premise argument when the negative evidence is instantiated by an exemplar from the same subcategory.

Method

Participants. 163 students from the University of Leuven, Belgium, participated in the study. Participants received course credits in return for their participation.

Design. In a between-subjects design, participants were presented with single positive premise or mixed positive and negative premise arguments with four possible hierarchically nested conclusions. The task was to select the conclusion with the broadest tenable scope given the premises. The mixed premise arguments contained either negative evidence from the same or a contrasting subcategory as the positive premise or from a different superordinate category entirely. Participants were thus allocated to one of four conditions, one of which evaluated the single positive premise arguments and the remaining three the mixed premise arguments. Participants chose between four possible conclusions: the exemplar contained in the positive premise (e.g., Shostakovich); a salient subcategory that the exemplar belongs to (e.g., classical music), its basic level category (e.g., music) or its superordinate category (e.g., sound). The responses hence reflect the breath of the generalization elicited by the given premises.

Materials. Ten target items and 30 filler items were created. All items were arguments consisting of either a single positive premise or a positive and a negative premise with four possible conclusions. The conclusions were hierarchically nested going from the exemplar presented in the positive premise up to its superordinate category. The premises and the conclusion of each argument contained exemplars from a single basic level category (e.g., insects, fruit, wines, car companies). Our three conditions for the double premise arguments varied in the type of negative evidence that were included in the target items. The negative evidence either came from the same subcategory (e.g., Hayden) as the positive premise or a contrasting subcategory (e.g., AC/DC) or a different superordinate category (e.g., the sound of a falling screwdriver). The properties used in the arguments were realistic characteristics that participants were likely to have very little knowledge about (e.g., produce ocytoncine; have mitochondrion in their cells; create a conversion current). In addition to our target items, we used 30 filler items across the four conditions that mimicked each of the conditions to reduce an effect of the ratio of the different item types across conditions. 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 groups of 25 and took no longer than 5 minutes to complete the task.

Results & Discussion

Figure 1 shows the proportion of responses across the four possible conclusions for each of the four conditions. Single premise arguments showed a monotonic decline across the conclusions from the exemplar to the superordinate. For arguments with negative evidence from the same subcategory, most responses shifted to the exemplar level. Those arguments with negative evidence from a contrasting subcategory showed most responses divided between the exemplar and subcategory level. In contrast negative evidence at the broadest level lead to evenly spread responses across all level bar the superordinate level.

In order to confirm the reliability of these shifts in response proportions across the four conclusions a 3 × 4 mixed factorial analysis of variance (ANOVA) across participants (F_1) and items (F_2) was carried out (Clark, 1973). The fourth response option was omitted from the analyses to avoid violations of the independence assumption of ANOVA. The analysis revealed a two-way interaction between Responses and Condition ($minF'(6,135) = 8.2$, $p < .001$). Broken down by Response, there was a significant shift of the proportion of responses for each of the response options across the four conditions except the superordinate level conclusion (Exemplar: $minF'(3,81) = 10.0$, $p < .001$;

Subcategory: *minF'*(3,53) = 3.7, $p < .05$; Category: *minF'*(3,85) = 16.6, $p < .001$; Superordinate: F_1(3, 162) = 3.0, $p < .05$, F_2(3, 27) = 3.8, $p < .05$, *minF'*(3,109) = 1.7, $p = .176$).

These overall shifts confirm that pitching the negative evidence at different levels of hierarchy changes the scope of people's generalization. However they are not sufficient to claim that the mechanism of redistributing the probability can explain a rise in argument strength observed in Heussen et al. (2011). In order to do that we need to look more closely at the exemplar and subcategory level between the single positive condition and both the negative evidence from the same and contrasting subcategory condition. Starting with a comparison of the single with the mixed premise argument containing negative evidence from a contrasting subcategory, the results revealed a significant increase in probability at the subcategory level both across participants (t(79) = 2.2, $p < .05$) and items (t(9) = 2.3, $p < .05$). As a result, all three remaining response options showed a decrease in probability, however this decrease was only significant for the response at category level (Participants: t(79) = 2.7, $p < .01$; Items: t(9) = 3.1, $p < .05$). These results show that by redistributing the probabilities across an exhaustive hypothesis space a significant increase in the generalization to the subcategory level can be achieved.

For negative evidence that lowers argument strength and is hence in line with the Monotonicity Principle, we look at the difference between the single premise and mixed premise argument with negative evidence from the same subcategory. Here we would expect a significant shift in probability from the subcategory to the exemplar level and hence a decrease at the former and an increase at the latter. In a planned comparison across participants (t(78) = 5.3, $p < .01$) and items (t(9) = 3.9, $p < .01$) we found a significant increase in proportions at the exemplar level. All remaining response options show a decrease with the largest difference at the subcategory level (M_{diff} = .18, SE_{diff} = .04; Participants: t(79) = 4.2, $p < .001$; Items: t(9) = 2.2, $p < .05$) followed by the category level (M_{diff} = .09, SE_{diff} = .02; Participants: t(79) = 4.2, $p < .001$; Items: t(9) = 2.8, $p < .05$) and the superordinate level (M_{diff} = .03, SE_{diff} = .01; Participants: t(79) = 2.0, $p < .05$; Items: t(9) = 2.3, $p < .05$).

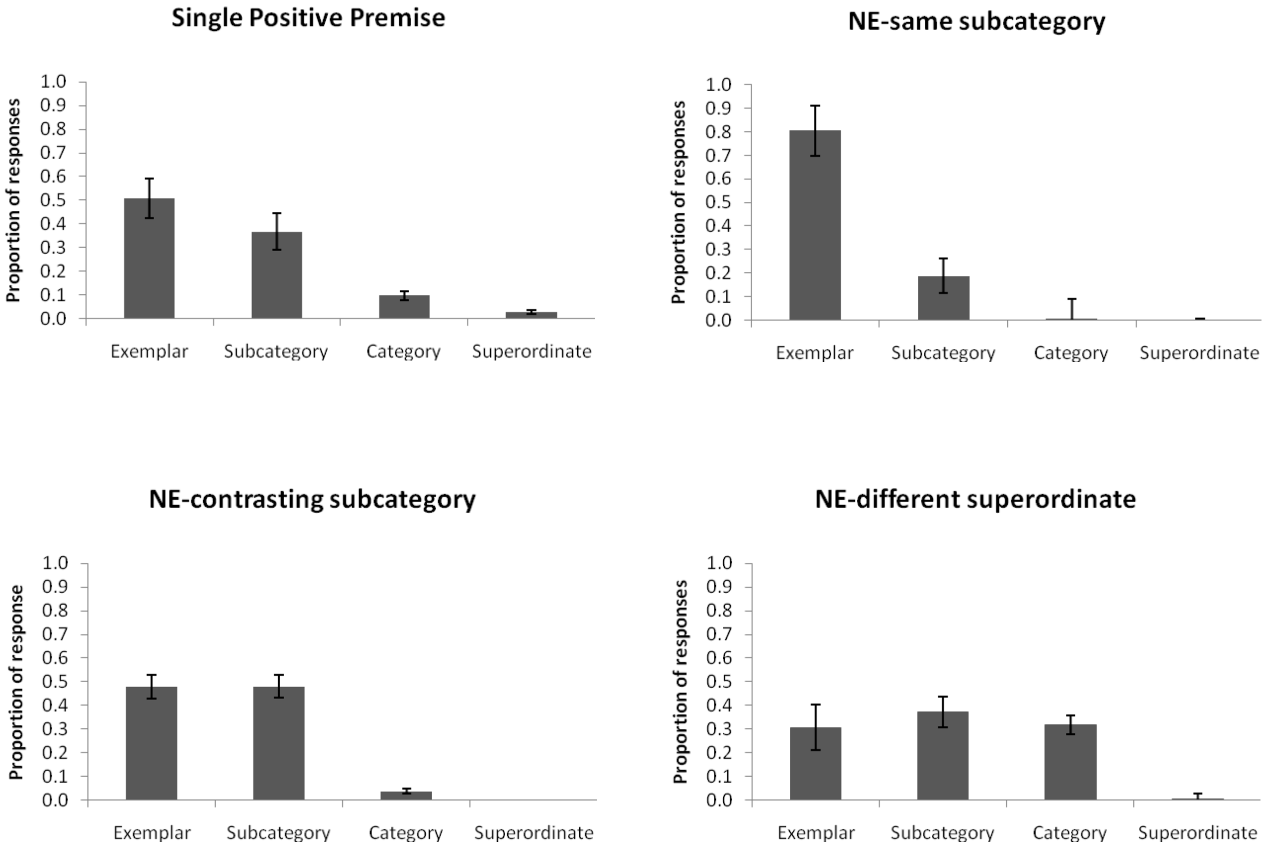

Figure 1: Average proportion of responses for selecting one of the four response options as a function of condition

General Discussion

In most cases, encountering negative evidence for a generalization lowers one's credence in that generalization. Contrary to this intuition however, Heussen et al. (2011) have found evidence that, in principle, it is possible to raise argument strength using negative evidence. In their study, presenting people with negative evidence from a subcategory that was contrasting both the positive premise and the conclusion led people to judge arguments containing negative evidence as stronger than single positive premise arguments. This implies that the range of the influence of negative evidence extends over the no effect point to a small but positive effect.

This poses two challenges for models of induction. First, models of induction to date use the 'sign' of the evidence to determine the direction of the influence (e.g., Blok, Medin & Osherson, 2007; Kemp & Tenenbaum, 2009). This implies that negative evidence can never raise argument strength. Hence models require a mechanism by which negative evidence can lead to an increase in argument strength. Second, even with a mechanism that could handle such an increase in argument strength, models would still require an a priori way to determine whether the negative evidence is of the kind that lowers argument strength or raises it.

Here we have only tried to address the first of the two challenges. How can argument strength increase from a single positive premise argument—Shostakovich elicits alpha waves, therefore Bach elicits alpha waves—to a double premise argument containing negative evidence—Shostakovich elicits alpha waves, the music of AC/DC does not elicit alpha waves, therefore Bach elicits alpha waves? The present results show that when people are faced with an exhaustive set of hypotheses—or in this case possible conclusions—it is possible to shift their preference to a particular conclusion by introducing negative evidence. More specifically, in comparison to the single premise argument, negative evidence from a contrasting subcategory to the positive evidence significantly increased people's preference for the generalization to the subcategory level. In contrast, negative evidence from the same subcategory led participants to shift their preference away from the subcategory level to the exemplar level generalization.

How do these results explain the increase in argument strength in the Shostakovich – AC/CD – Bach case? The idea is that when participants encounter the single premise argument, they develop a set of hypotheses and judge the believability of each one. These are assumed to be similar to the one's collected in the single positive premise condition. If confronted with the fact that the music of AC/DC does not cause alpha waves then participants think the generalization to the subcategory—in this case the one that Bach belongs to—is more likely than before, as evidenced by the condition with negative evidence from a contrasting subcategory.

What assumptions have to be met for this account to hold? First, people have to generate a list of possible hypotheses in response to the single premise. Second, at the time of encountering the negative evidence, that list must—at least in their mind—be exhaustive. Third, they have to assign a probability distribution across the set of hypotheses. The first and the third assumption are far from being controversial. People are easily able to generate a range of hypothesis and are even more sensitive to their accuracy than when evaluating other people's hypotheses (e.g., Dougherty & Hunter, 2003; Koehler, 1994). Furthermore the idea that people use probabilities to represent their fine-grained beliefs is common place in psychological models of induction (e.g., Heit, 1998; Kemp & Tenenbaum, 2009). Whether people's generated hypotheses constitute an exhaustive list in their minds is questionable. A humble person would surely admit the possibility that there are hypotheses that she has not yet considered. The solution for this would be to grant a bin category of hypotheses that have not yet been considered. A problem with that though is how much credence should one attribute to that category in comparison to the other hypotheses. In some circumstances, when people for instance have very little knowledge about a particular subject area, they may not be confident about their hypotheses and hence attribute a lot of credence to such a bin category. In those cases, it would be difficult for negative evidence to contradict the complete bin category and thereby raise the probability for the other hypotheses. Thus we might not expect to be able to increase people's judgment about argument strength in those cases. However in cases where people have some level of knowledge that enables them to have some confidence in their own hypotheses, it would seem odd to attribute a larger amount of credence to the bin category at the expense of the generated hypotheses. Thus in normal circumstances granting a bin category for hypotheses, would not constitute a problem and hence makes the assumption of an exhaustive hypothesis space rather tenable.

What are the implications of these finding for models of induction? For models of induction, these findings are good news because they provide an easy to implement mechanism to meet one of the challenges posed by negative evidence that raises argument strength (Heussen et al., 2011). The only thing that models of induction need to assume to implement this mechanism is a probability distribution across a set of generated hypotheses to represent people's relative credence in the truth or strength of those hypotheses. Assuming this mechanism, makes it at least in principle possible to raise argument strength with negative evidence. However, it does not address the question of when a raise actually occurs? How do models of induction distinguish between negative evidence that is detrimental to argument strength and negative evidence that raises argument strength?

Based on the relevance theory of induction, we would here like to propose a tentative mechanism of what makes negative evidence relevant either to a negative or a positive effect? Although not formally specified the relevance theory of induction (Medin et al., 2003) provides at least a framework for such a mechanism. The underlying idea is that distinctive properties of the premises bring to light relevant dimensions for induction. These dimensions are then either reinforced (in case of a match) or undermined (in case of a mismatch) by comparing the premises with the conclusion. If people find a relevant dimension for induction (e.g., classical music) that is common to the positive premise and the conclusion, negative evidence can either undermine or reinforce the validity of the dimension. The validity of a dimension is undermined when negative evidence shares that dimension with the positive premise and the conclusion (e.g., Haydn doesn't elicit alpha waves, thus classical music cannot be the basis for induction) and reinforced when it does not share that dimension (e.g., AC/DC does not elicit alpha waves but it is also not classical music). Whether negative evidence that reinforces a dimension is then considered relevant enough to increase argument strength depends on whether the negative evidence increases the salience of the dimension sufficiently above what it would have been without the negative evidence. In other words, the likelihood of a generalization from Shostakovich to Bach will increase with the introduction of negative evidence, if the negative evidence raises the salience of classical music as a basis for induction. How might that happen? The relevance approach suggest that both the level of effort necessary to process an input and the effect that such an input has, affect the relevance of such an input (Sperber & Wilson, 1995). Hence, if the negative evidence lowers the effort necessary to draw out the dimension used for induction, then inductive strength may increase. Furthermore, inductive strength may increase, if the introduction of negative evidence highlights a particular dimension that brings about a stronger effect than a dimension that had been considered before the introduction of negative evidence.

Acknowledgments

The work was supported in part by an F+ Fellowship within the framework of international mobility awarded to the first author by KU Leuven under the supervision of Gert Storms. The authors would like to thank Steven Verheyen and James Hampton for helpful input to the project.

References

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

Dougherty, M. R. P., & Hunter, J.E. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica, 113*, 263-282.

Feeney, A., & Heit, E. (2007). *Inductive reasoning*. Cambridge University Press.

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

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

Heussen, D. & Hampton, J. A. (2009). Counterexamples in category-based property induction. Poster presented at the 50th Annual Meeting of the Psychonomics Society.

Heussen, D., Voorspoels, W. & Storms, G. (2010). Can similarity-based models of induction handle negative evidence. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 2033-2038)*. Austin, TX: Cognitive Science Society.

Heussen, D., Voorspoels, W., Verheyen, S., Storms, G., & Hampton, J.A. (2011). Raising argument strength using negative evidence: A constraint on models of induction. *Manuscript submitted for publication*.

Houtz, J. C., Moore, J. W., & Davis, J. K. (1973). Effects of different types of positive and negative instances in learning "nondimensioned" concepts. *Journal of Educational Psychology, 64*, 206–211.

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.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of experimental psychology: Learning, Memory and Cognition, 20*, 461-469.

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

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review, 10*, 517-532.

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

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

Russell, B. (1912). *Problems of Philosophy*. Oxford University Press (1972).

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

Smith, E.E., Shafir, E., & Osherson, D.N. 1993. Similarity, plausibility, and judgments of probability. *Cognition, 49*, 2, 67-96.

Sperber, D. & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.) Oxford: Blackwell.