RUNNING HEAD: A constraint on models of induction

Raising argument strength using negative evidence: A constraint on models of induction

Daniel Heussen
Wouter Voorspoels
Steven Verheyen
Gert Storms
University of Leuven

James A. Hampton
City University, London

word count: 7000

Address for correspondence:
Daniel Heussen
Psychology Department
University of Leuven
Tiensestraat 102
3000 Leuven, Belgium
daniel.heussen@psy.kuleuven.be

Raising argument strength using negative evidence

Abstract

Intuitively and according to general predictions of similarity-based theories of induction, relevant evidence raises argument strength when it is positive and lowers it when it's negative. Three experiments test the hypothesis that argument strength can actually increase when encountering negative evidence. Participants made forced choice judgments between (Experiment 1 & 2) or sequentially evaluated (Experiment 3) single positive (e.g., Shostakovich's music causes alpha waves in the brain, therefore Bach's music causes alpha waves in the brain) and double mixed premise arguments (e.g., Shostakovich's music causes alpha waves in the brain, X's music DOES NOT, therefore Bach's music causes alpha waves in the brain) where the second premise (i.e., X) was either from the same subcategory as the first premise and the conclusion (e.g., Haydn) or from a different subcategory (e.g., AC/DC). Negative evidence lowered credence when it applied to an item from the same subcategory and rose when it was applied to a different subcategory. The results constitute a new constraint on models of induction.

Keywords: Induction, Negative evidence, Categories, Models of induction

Introduction

Everyday reasoning consists for the most part of inductive inference, where induction in its broadest sense constitutes an inference to an uncertain conclusion. One strategy to make such an inference is to use past experiences—my car has always started, so it's reliable and will start today. Another strategy is to use category membership—my car is a German car, so it's reliable and will start today. Research on induction in psychology has predominately focused on the latter of the two. In a typical experimental setup, people are told that one or several categories have a particular property and are asked to extend that property to other categories within a given domain. Participants, for instance, 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). Numerous phenomena relating to category-based property induction have been documented (e.g., Hampton & Cannon, 2004; Heit, 2000, for a summary; Heit, Hahn, & Feeney, 2004; Rips, 1975) and various models have been proposed to account for people's judgments of inductive arguments like these (e.g., Blok, Medin, & Osherson, 2007; Heit, 1998; Kemp & Tenenbaum, 2009; Medin, Coley, Storms, & Hayes, 2003; Osherson, et al., 1990; Sloman, 1993).

In the majority of cases, experimental work and modeling efforts have focused on arguments involving positive evidence; premises that state that some entity possesses the to-be-projected property. In reality though, we do not only receive positive evidence for our hypotheses. We are often confronted with negative evidence; evidence that states that some entity of the same or a similar category DOES NOT possess the to-be-projected property. For instance, in evaluating whether *Bobcats* use serotonin as neurotransmitter we might find out that *Tigers* do but *Cougars* do not. How do we integrate the negative evidence and how does it influence our judgment about *Bobcats*?

In the present paper we are interested in the influence of negative evidence on argument strength. More precisely, we are interested in the direction of this influence. Contrary to intuition and the general predictions of similarity-based theories of induction, we present evidence that, under certain circumstances, negative evidence can actually facilitate the generalization of a property. In what follows, we will identify and discuss in more detail the general assumption that negative evidence has a negative effect on argument strength. Then we will present three experiments that undermine the universal validity of this assumption. In the General Discussion, we address in more detail the implications of these findings for well-known models of induction.

Monotonicity in inductive reasoning

In absence of other information, it is generally assumed that generalization relies on similarity (Blok, Medin, & Osherson, 2007; Osherson et al., 1990; Rips, 1975; Shepard, 1987; Sloman, 1993; Smith, Shafir & Osherson, 1993). In the context of category-based property induction, similarity between premise and conclusion categories determines the influence that a premise has on argument strength: The more similar the premise and conclusion categories are, the greater the influence of the evidence on the conclusion (e.g., Rips, 1975). In evaluating whether *Bobcats* have property x, knowing that *Tigers* do seems to provide stronger evidence than discovering that *Penguins* do. Moreover, this relation between similarity and arguments has been shown to hold for additional premises—discovering that not only *Tigers* have property x, but also *Lions* and *Cats*, will further strengthen one's belief that indeed *Bobcats* too must have the property. The tendency that accumulating further positive premises is positively related to argument strength has been referred to as monotonicity in inductive reasoning, and is supported by many empirical results (e.g., Feeney, 2007; McDonald, Samuels, & Rispoli, 1996; Osherson et al, 1990, 1991; for summaries see, Hayes, Heit & Swendsen, 2010; Heit, 2000).

Now, what can we expect when negative evidence comes into play? It is important to realize that there is no reason to consider not-having-a-property as essentially different from having-a-property[1]. In other words, discovering that *Lions* do not have property x provides a reasonable basis to conclude that *Bobcats* do not have x, analogously to having a property. Obviously, the conclusion that *Bobcats* do not have x, is inversely related to the conclusion that *Bobcats* have x. Hence *Bobcats* having x becomes less likely in the light of *Lions not* having x. Following a similarity-based approach, negative evidence therefore decreases the argument strength regarding a positive conclusion and a tendency to monotonicity can be expected when adding more negative evidence. While research on negative evidence is few and far between, results support this general tendency of monotonicity (e.g., Blok et al., 2007; Heussen & Hampton, 2011; Osherson et al, 1991).

In short, evidence seems to raise argument strength if it is positive evidence (i.e., when it states that some entity has the property) and to lower it if it is negative evidence (i.e., when it states that some entity does NOT have the property), with similarity determining the size of the change in argument strength. Hence, argument strength does not seem to move in the opposite direction from the "sign" of the evidence. In line with previous research, we call this general tendency the Monotonicity Principle about the influence of evidence on argument strength. To the extent that theories of inductive reasoning rely on similarity, models of induction endorse the Monotonicty principle as presented here (e.g., Blok, Medin, & Osherson, 2007; Osherson et al., 1990; Sloman, 1993). The models can however have other mechanisms that may explain violations of the principle (e.g., Osherson et al., 1990); this will be discussed in more detail in the General Discussion.

In search of evidence against monotonicity

On the side of positive evidence, there are some notable violations to the principle (Medin et al, 2003; Osherson, et al., 1990). These violations of monotonicity seems to be based on using additional evidence that comes from a category that is different from the one that includes both the original premise and conclusion categories. For example, Osherson et al. (1990) reported that a greater proportion of people preferred an argument from *Fly* to *Bee* than from *Fly* and *Orangutan* to *Bee*. Adding positive evidence from a category (mammals) different to that of the conclusion and first premise (insects), seems to elicit a drop in argument strength, thus disconfirming the monotonocity principle for positive evidence. A related phenomenon has been reported by Medin et al. (2003), who showed that people judged an argument from *Polar bear* to *Penguin* as stronger than the argument from *Polar bear* and *Brown bear* to *Penguin*. Again, adding positive evidence decreases the argument strength, contrary to what the monotonicity principle would predict. However, in this case it is the addition of positive evidence from a more specific category (bears) than the category including the first premise and the conclusion (animals).

The symmetry between positive (i.e., having a property) and negative evidence (i.e., not having a property) suggests that similar violations of the monotonicity principle should be found for negative evidence. Moreover, there is evidence from developmental studies that suggests that negative evidence or more precisely contrastive information can have a beneficial effect on generalizations, at least among children (Kalish & Lawson, 2007; Waxman, Lynch, Casey & Baer, 1997). However due to departure from a traditional paradigm in (i) using *individuals* rather than whole classes, (ii) the use of implicit rather than explicit negation and (iii) less than

unequivocal results in the adult sample, it is unclear to what extent these findings generalize to a standard category-based property induction paradigm within an adult population.

The present study aims to establish non-monotonicity effects in a standard category-based induction task when adding negative evidence to an argument. We do not aim to offer a definitive theoretical account of how or why inductive strength can rise with negative evidence, nor are we making a universal claim about the circumstance or conditions in which the effect occurs. Rather, our aim is to make an existential claim that there are cases in which argument strength can rise by finding out about some negative evidence.

Across three experiments we used a category-based property induction paradigm with blank properties—properties that participants are likely to have very little knowledge about. Participants made forced choice judgments between (Experiment 1 & 2) or sequentially evaluated (Experiment 3) single positive (e.g., Given that freight ships create conversion currents, how likely is it that cruise ships do so) and double mixed premise arguments (e.g., Given that freight ships create conversion currents, and that hovercraft ships DO NOT, how likely is it that cruise ships do so).

To elicit the effect, we presented participants with exemplars from the same subcategory for the positive evidence and the conclusion and used another subcategory for the negative evidence. Subcategories are here not to be understood as rigid classes within a fixed taxonomic hierarchy. Arctic animals could just as well be a subcategory of animals as felines or canines. The use of evidence from different subcategories within a common superordinate category was intended to achieve a demarcation of relevant dimensions or criteria for induction.

Our underlying hypothesis for the three experiments was that a preference for arguments containing negative evidence (i.e., a preference for mixed over single premise arguments) will be more likely when negative evidence is instantiated by an item from a contrasting subcategory

than those of the positive premise and the conclusion. In contrast, if the negative evidence is instantiated by an item from the same subcategory the usual negative impact on argument strength is expected.

Experiment 1

In Experiment 1 participants were asked to choose the stronger of two arguments, one with a single positive premise and the identical argument with an additional negative premise.

	Shostakovich elicits alpha waves.
Shostakovich elicits alpha waves.	Music of AC/DC does not.
Bach elicits alpha waves.	Bach elicits alpha waves.

If the Monotonicity Principle is true, adding a negative premise to an argument with a single positive premise should in principle not increase argument strength. Hence, participants should always choose the single positive premise. Any deviation from preferring the single premise argument indicates a violation of the Monotonicity Principle.

Method

Participants. Participants were 32 first year undergraduate students at the University of Leuven who each completed a booklet for course credit.

Design. In a repeated measures design, participants made forced choice judgments about a list of 30 pairs of arguments. Participants were instructed to assume that the premises of the arguments are stating facts and asked to make a forced choice for the argument whose premises provide better reasons to belief the conclusion. Each pair consisted of a single and a mixed premise argument. The mixed premise argument was identical to the single premise argument with the exception of an added negative premise. Half of the pairs were target pairs and half were control pairs. Two random orders of items were used.

Materials. The premises and the conclusion of each argument contained exemplars from a single category (e.g., insects, fruit, wines, car companies). For the target items, the positive premise and the conclusion were from one loosely defined subcategory (e.g., flying insects, tropical fruit, European wines, German car companies), whereas the negative premise was an exemplar from a contrasting subcategory (e.g., crawling insects, Northern European fruit, New World wines, Italian car companies). The negative evidence in the control items came from the same subcategory as for the positive premise and conclusion. The selection of items from loosely defined subcategories was based on the first two authors' intuitions[2]. 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). A list of the items used in all three experiments can be found in the Appendix.

Procedure. Students participated in groups and completed the questionnaire as part of a series of tasks. The task took no more than 5 minutes.

Results

Figure 1 shows the average proportion of responses across 15 target and 15 control items that showed a preference for the mixed premise argument containing negative evidence over the single positive premise argument.

INSERT FIGURE 1 ABOUT HERE

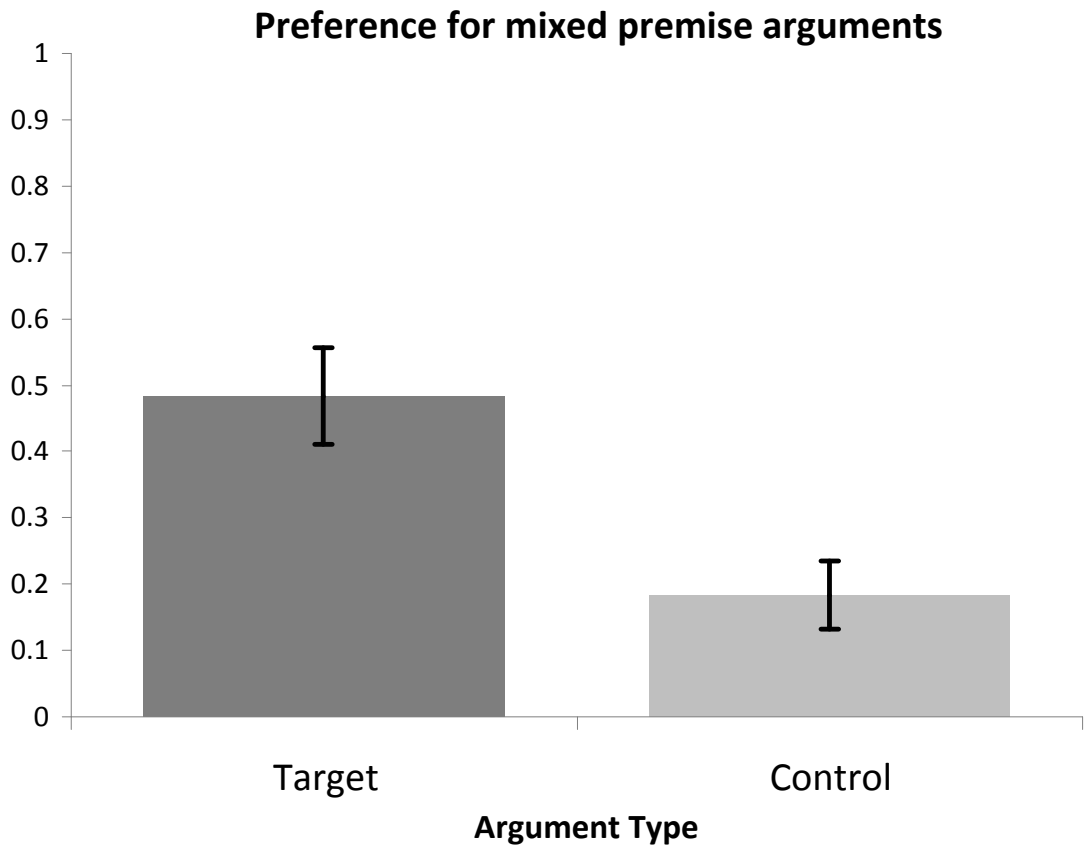


Figure 1. Average proportion of responses selecting the mixed premise argument for target and control items. Error bars constitute 95% confidence intervals calculated on the participant data.

An independent sample t-test on the average proportion of participants ($t(28) = 6.4$, $p < .001$) and a paired-sample t-test on the average proportion of items ($t(31) = 6.5$, $p < .001$) showed a significant difference (min $F'(1, 59) = 20.8$, $p < .01$) between our target and control items in response preference for arguments involving negative evidence with both proportions being significantly higher than zero (Control: min $F'(1, 43) = 28.9$, $p < .01$; Target: min$F'(1, 35) = 74.5$, $p < .01$). Participants were two and a half times more likely to select the argument containing negative evidence as stronger among target items than among control items.

Preference for negative evidence arguments rose from just under 20% in the control items to nearly 50% in the target items.

Discussion

Experiment 1 demonstrated in a forced choice paradigm that participants can be tempted into preferring an argument that contains negative evidence to those with just a single positive premise. This preference was significantly greater when the negative evidence was instantiated by an item from a different rather than the same subcategory as that of the positive premise and the conclusion. Experiment 1 therefore provides evidence that human reasoning in some cases violates the Monotonicity Principle about the influence of negative evidence. Two important remarks are in place here, the first concerning the control items, the second relating to the target items. We will discuss these in turn.

Responses for the control items were not at normative levels with nearly 20% of responses on average across participants violating the Monotonicity Principle. A plausible explanation of this relates to the design characteristics of the study. Participants were asked to make 30 forced choice judgments. If the Monotonicity Principle about negative evidence is correct, then participants would have had to consistently choose the single premise argument, resulting in no legitimate variation in response pattern. Thus one might argue that, through the design of the study, participants were somewhat coerced into indicating a preference for the mixed premise argument at least for a few of the control items. Experiment 2 will address this possibility.

Nevertheless, the responses to the control items can be seen as the baseline for our target items even if they are not located at the normative value. If the baseline preference for arguments containing negative evidence in an experimental setup like this is around 20%, then our

participants still showed a significantly greater preference for mixed premise arguments among target items than among control items. Hence they were sensitive to the similarity between the positive and negative premise items in making their choice.

A second point of discussion relates to the average proportion of responses endorsing the negative evidence arguments of the target items, amounting to nearly 50%. One could argue that this proportion reflects responses at chance level. It is possible that for the target items, participants perceived the strength of single and mixed premise arguments as identical, hence making a choice based on a coin flip. If this is the case, the present finding does not support the conclusion that adding negative evidence can increase argument strength, but merely shows that there are cases in which negative evidence does not lower argument strength. Note that, following the Monotonicity Principle, this implies that the negative premise concerns a category that is irrelevant—otherwise it would lower argument strength.

We believe that this is not the case. In order to construct the target items, we chose categories that were explicitly similar in a very salient respect (belonging to the same superordinate category, e.g., music). Consequently, these categories bear relevance to the arguments in question, and following the Monotonicity Principle, should always lower argument strength. This prediction is contradicted by the data. Participants were clearly tempted to choose the argument that included negative evidence at a rate significantly above zero (the prediction of the Monotonicity Principle), and significantly more than the control items (which form an empirical baseline). The rise in argument strength provided by the addition of negative evidence was not however extremely large, making the choice blatantly clear for the participants. But against a low background expectation of a preference for arguments with negative evidence, the level seen was nevertheless high.

The following two studies aimed to back this claim by contrasting relevant with irrelevant negative evidence (Experiment 2) and by explicitly demonstrating an increase in rated argument strength from single to mixed premise arguments within participants (Experiment 3).

Experiment 2

Experiment 1 used a within-subjects design with target and control condition being instantiated by different items. In Experiment 2, we replicated the study using a between subjects design in order to test whether the manipulation to elicit the effect is robust within items. As before, participants were asked to choose between a single premise argument and a mixed premise argument (containing a negative premise). Different groups of participants were presented with the same positive premise and conclusion but different negative evidence premises that either came from a different subcategory as the positive premise and the conclusion (Target), or from the same subcategory (Control). In addition, we added a third condition, in which the additional evidence was negative but irrelevant (Irrelevant):

Shostakovich's music elicits alpha waves.
A falling rock does not elicit alpha waves.
Bach's music elicits alpha waves.

The Irrelevant condition allows us to test whether responses to the target condition in Experiment 1 constituted chance level responding. Chance level responding in the target condition implies that the negative evidence is irrelevant for the conclusion. If the preference for arguments containing negative evidence in the irrelevant condition of Experiment 2 can be shown to be significantly lower than in the target condition, then responses to the target condition—even if they are at .5—can not constitute chance level responding and hence indicate a rise in argument strength.

Thirty filler items were also constructed to elicit all possible response patterns. Fillers remained the same across the three conditions. These two additions counter the potential objections we raised in the discussion above.

Method

Participants. Participants were 121 first year undergraduate students at the University of Leuven who each completed a booklet for course credit.

Design & Materials. The task was identical to Experiment 1. Participants were asked to judge which of two arguments (a single and a mixed premise) provides better reasons to belief the conclusion. A between subjects design was used in which 10 of the target items in Experiment 1 (target items that showed the strongest effect) were presented in three different conditions—target, control and irrelevant. Participants were randomly allocated to one of the three conditions with roughly 40 participants in each condition. As previously, the target and the control condition contained negative evidence from a different or the same subcategory as the positive premise and the conclusion, respectively. In addition, an 'irrelevant' condition presented negative evidence in the form of an exemplar from a different superordinate category than that of the positive premise and the conclusion.

The same thirty filler items were used across the three conditions. These consisted of 10 purely positive argument pairs (e.g., Lions have enzyme x, tigers have enzyme x. How likely is it that cheetahs have enzyme x?) that should clearly elicit a preference for the mixed premise argument in order to provide legitimate variation in response choices and 20 argument pairs with negative evidence that should elicit a preference for the single premise argument in form similar to our control items.

Procedure. Students participated in groups and completed the questionnaire as part of a series of tasks. The task took no more than 8 minutes.

Results & Discussion

In line with Experiment 1, *Figure 2* shows the average proportion of responses that indicate a preference for mixed premise arguments containing negative evidence over single positive premise arguments.

INSERT FIGURE 2 ABOUT HERE

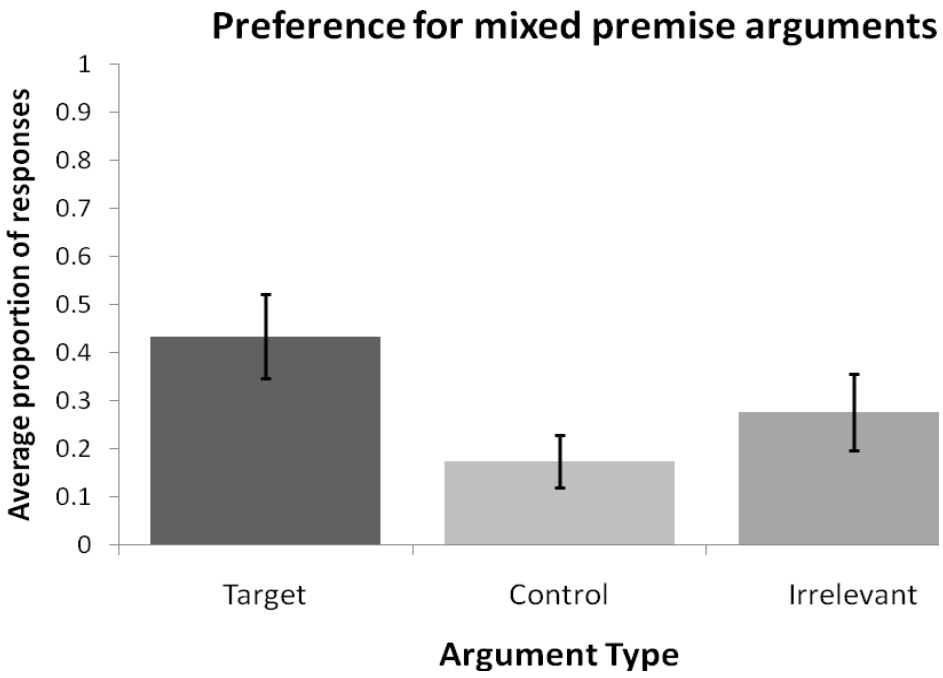


Figure 2. Average proportion of responses selecting the mixed premise argument for target, control and irrelevant items. Error bars constitute 95% confidence intervals calculated on the participant data.

A one-way ANOVA across participants and items showed a significant difference in preference for mixed premise arguments between conditions, min $F'(2, 89) = 7.53$, $p < .001$. Planned pair-wise comparison for participants and items showed a significant difference between the target and control condition (min $F'(1, 42) = 14.35$, $p < .001$) and between target and irrelevant (min $F'(1, 65) = 4.87$, $p < .05$), but not between irrelevant and control (min $F'(1, 48) = 2.79$, $p = .102$).

Experiment 2, thus, replicated the findings of Experiment 1 in a between-subjects design providing evidence that participants violate the Monotonicity Principle. The same single premise arguments were either preferred or rejected depending on the additional negative evidence. People showed a greater preference for arguments containing negative evidence when the evidence came from a different rather than the same subcategory as the positive premise and the conclusion. Furthermore, using irrelevant negative evidence from an unrelated superordinate category resulted in responses closer to the control than to the target condition indicating that irrelevant evidence in a forced choice paradigm leads to a preference for single premise arguments and not coin-flip responding. If the 50% choice for the target arguments in Experiment 1 had simply reflected a view that the negative premise was irrelevant, then we should have observed a similar level of choice for the irrelevant condition in Experiment 2. The significant difference between these two conditions therefore rules out this account.

Experiment 3

In order to explore this violation of the Monotonicity Principle further, Experiment 3 was designed to replicate the effect in a new paradigm. We changed our procedure from a forced choice to a sequential judgment task. Participants first evaluated the single premise argument, then received the negative premise and evaluated the mixed premise argument.

Our predictions were an increase in argument strength from single to mixed premise arguments for the target items but not for the control items. The control items should reflect a clear drop in argument strength in line with the Monotonicity Principle. Note that in the present experiment, we measured the effect within a person and an item, the strongest test for the claim that negative evidence can indeed raise argument strength. In addition people were able to express the view that the strength was unaffected by giving the same rating to each argument.

Method

Participants. Fourteen undergraduate students at the University of Leuven participated for course credit.

Design and Materials. A repeated measures design was used to measure the change in argument strength from single to mixed premise arguments. Participants were asked to judge both single and mixed premise arguments as well as whether argument strength had decreased, stayed unchanged or had increased between the first and second judgment. Judgments of argument strength were measured on an 11 point scale. Two random orders of items were used.

The arguments were identical to those used in Experiment 1. The arguments were imbedded in little vignettes describing the first premise as a well-established scientific or specialist fact and then asking participants to evaluate the conclusion. Participants then received a second piece of information again described as a well-established fact and were asked to evaluate whether this lowered, raised or left their judgment of the conclusion unchanged. Subsequently they gave a final judgment of the conclusion given both facts.

Procedure. Participants completed the booklets within an individual testing session.

Results & discussion

Figure 3 shows the average proportion of responses indicating a decrease, no change or increase in argument strength from the first to the second judgment for target and control items.

INSERT FIGURE 3 ABOUT HERE

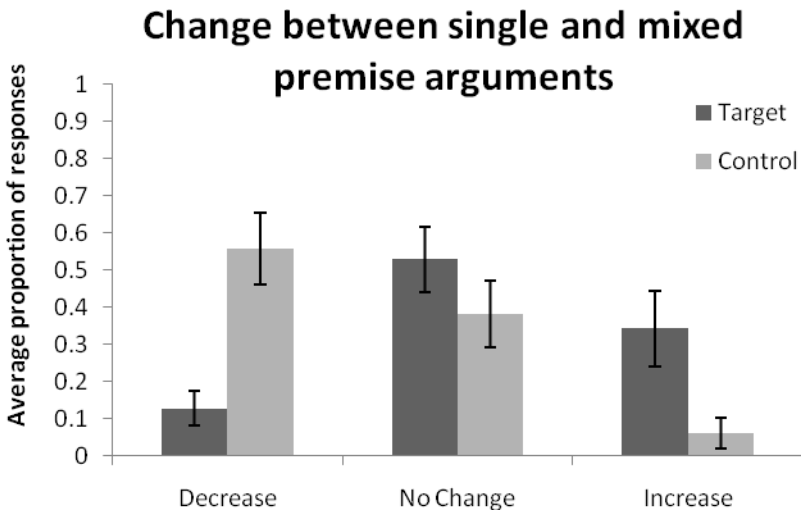


Figure 3. Average proportion of responses indicating a change in argument strength from the single positive (first judgment) to mixed premise arguments (second judgment) for target and control items. Error bars constitute 95% confidence intervals calculated on the participant data.

A 2 × 2 ANOVA[3] for participants and items revealed a significant interaction between the direction of change in people's judgment (i.e., Decrease, Increase) and Item Type, *min F'*(1,39) = 45.3, $p < .001$. For the target items a greater average proportion of responses across participants and items indicated an increase rather than a decrease in argument strength from the first to the second judgment (*min F'*(1,26) = 5.8, $p < .05$), whereas for the control items the reverse pattern was found (*min F'*(1,26) = 37.5, $p < .001$). Although for the target items the greatest proportion of items did not change in strength, indicating that negative evidence for those items was

perceived as irrelevant, the average proportion of responses indicating an increase was significantly larger for the target items than for the control items (Target: $M = 0.34$, $sd = 0.19$; Control: $M = 0.06$, $sd = 0.07$; $minF'(1,36) = 22.7$, $p < .001$).

This pattern of results was also reflected in the ratings of the single and mixed premise arguments. *Figure 4* shows the average argument strength of single positive and mixed premise arguments for target and control items.

INSERT FIGURE 4 ABOUT HERE

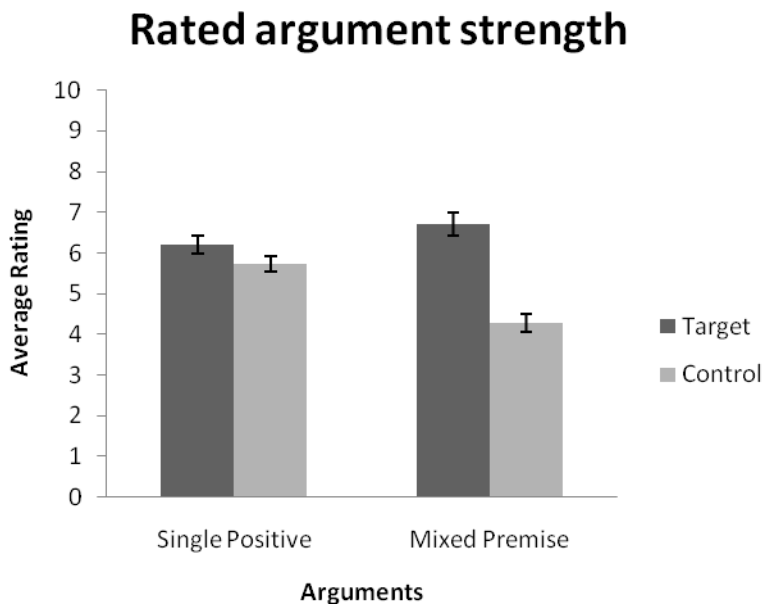


Figure 4. Average argument strength of single positive (first judgment) and mixed premise arguments (second judgment). Error bars constitute 95% confidence intervals calculated on the participant data.

A 2 × 2 ANOVA for participants and items revealed a significant interaction between the Type of Argument (levels: Single premise vs. Mixed premise arguments) and Item Type (levels: Target vs. Control items), $min\ F'(1,39) = 38.8$, $p < .001$. Target items showed a significant

increase in argument strength from single to mixed premise arguments (*min F'*(1,26) = 7.0, $p < .05$), whereas control items showed a significant decrease (*min F'*(1,25) = 33.9, $p < .001$). Among the single premise arguments, target and control items did not show a significant difference in argument strength (Item: $F(1,13) = 6.2$, $p < .05$; Participant: $F(1,28) = 2.7$, $p = 0.114$; *minF'*(1,41) = 1.9, $p = 0.178$). These results confirm our contention that it is possible to increase argument strength by introducing negative evidence.

Experiment 3 tested the Monotonicity Principle in a sequential evaluation task within items and participants and hence constitutes the most stringent of the three studies. Although the majority of people indicated no change in ratings of our target items between the first and the second judgment, a significant proportion did indicate an increase[4]. If the results of Experiment 1 & 2 constituted coin-flip responding on the grounds that the negative premise was irrelevant and did not change the argument, here we should have seen very few responses indicating the Increase option, and almost all responses indicating the No Change option. Experiment 3 does not seem to support this view. To the contrary, in line with the proportions, average ratings of the target items went up significantly from the single to the mixed premise arguments providing clear evidence for a rise in argument strength. In contrast, the control items showed a substantial decrease in argument strength.

General Discussion

Imagine you want to test the proposition that 'all ravens are black'. There are two related hypotheses that need to hold for the proposition to be true. One is 'if it's a raven, then it is black'. And the second is, 'if it's not black, then it's not a raven'. Intuitively, seeing something that is a raven and that is black confirms the hypothesis 'if it's a raven, then it is black'. But similarly seeing something that is not black and not a raven confirms the hypothesis 'if it's not black, then it's not a raven'. Hempel (1945) demonstrated the logical equivalence of these two hypotheses,

showing that evidence that supports one hypothesis logically provides support for the other. Hence, based on formal logic, encountering a 'white swan' is confirmatory evidence for the proposition 'all ravens are black'. But as Hempel argued and most readers would agree, this seems counterintuitive and hence poses a paradox between formal logic and intuition (Hempel, 1945).

Hempel's paradox depends on what we called the Monotonicity Principle of the influence of evidence. In the present paper, however, we have demonstrated that it is possible to raise argument strength by providing negative evidence. There are arguments in which figuratively speaking 'seeing a white swan' does provide evidence for the conclusion 'all ravens are black'. Although this finding does undermine the paradox at the side of intuition, as such it does not eliminate the paradox. According to logic, any non-black non-raven provides support for the conclusion 'all ravens are black' and that is clearly not the case. More modestly, our finding highlights the need for an account that is able to differentiate between negative evidence that lowers and negative evidence that raises credence in the conclusion of an argument. Such an account would provide a psychological demarcation of Hempel's paradox.

Models of induction and non-monotonicity

The present findings not only have implications for Hempel's paradox but raise an important question concerning existing models of induction, namely, how to incorporate the influence of negative evidence. To date, most models of induction have focused on the influence of positive evidence (e.g., Osherson et al., 1990; Rips, 1975; Sloman, 1993) with only two models providing an explicit formulation of the influence of negative evidence (Blok, Medin & Osherson, 2007; Kemp & Tenenbaum, 2009). In order to provide some comparison of the main models we'll make minimal assumptions on how negative evidence could be implemented. While an elaborate

presentation of the models and possible adaptations to account for our results falls outside the scope of this contribution, a model-based analysis of the present finding can lead to a better understanding of the present phenomenon.

Feature-based induction model

The feature-based induction model (FBIM: Sloman, 1993) derives the strength of a conclusion from the association between existing and novel features that is built from the evidence in the premises. Generalization therefore increases as similarity increases. While FBIM is not formulated for negative evidence, assuming the symmetry of having-a-property and not-having-a-property, the model's predictions are in line with the Monotonicity principle and hence cannot handle our findings.

SimProb

In the SimProb model (Blok, Medin & Osherson, 2007), a judgment of argument strength is conceptualized as a conditional probability judgment that relates the prior probability of a conclusion to the relevance of the evidence (by degree of similarity) and the degree to which the evidence is surprising (by prior probability of the evidence). The SimProb model explicitly allows for negative evidence premises, yet does not allow a rise in argument strength following the addition of a negative premise. According to the SimProb model, relevant evidence will raise argument strength if it is positive and lower argument strength if it is negative. As such, the model endorses the Monotonicity principle and hence cannot handle the present findings.

Similarity-Coverage model

The well-known similarity-coverage model (SimCov: Osherson et al., 1990) relates the strength of an argument to two similarity-based components. First, the similarity between the premise and conclusion categories plays an important role. In its present formulation the model does not incorporate negative evidence, however assuming a symmetry between having-a-property and

not-having-a-property leads to model predictions that are necessarily in line with the Monotonicity principle. In other words, the similarity term cannot accommodate a raise in argument strength due to negative evidence[5].

The second component that can influence argument strength according to the SimCov model is the coverage term, which reflects the extent to which an immediately relevant superordinate category is "covered" by the premise categories. Coverage is operationalized by computing the average maximum similarity of the premise categories to other members of the nearest superordinate category that includes premises and conclusion. Interestingly, the coverage component is crucial in explaining non-monotonicity effects for strictly positive premises (Osherson et al., 1990), due to changes in the relevant superordinate category for the argument (see Introduction).

If we assume that in the single premise arguments (e.g., Shostakovich → Bach) people rely on the broader basic-level category (e.g., MUSIC) as the immediate superordinate in the coverage term, the addition of the negative premise (e.g, AC/DC does not have the property) forces a change of the-to-be-covered category to a lower subcategory (e.g., CLASSICAL MUSIC). The latter, more specific category, is clearly better covered by the positive premise category. Therefore, adding the negative evidence may actually raise argument strength because the positive premise provides greater coverage for the more specific subcategory (i.e., CLASSICAL MUSIC) compared to the more general basic level category (i.e., MUSIC). Presented like this, the observed effect is a negative evidence variant of the non-monotonicity effect described by Osherson et al.

Note however, that this assumes that the coverage term only covers the positive premise and conclusion categories and not the negative premise categories. Furthermore, according to SimCov, people turn to the most specific superordinate category that includes both premise and

conclusion categories. It is therefore unclear why in the case of negative evidence the basic level (e.g., MUSIC) should be considered the to-be-covered category for single positive premise arguments. Moreover, even if changing the to-be-covered category was the mechanism to deal with negative evidence that raises argument strength, it is then unclear how the SimCov model would handle our control items, in which the negative evidence that is at the same hierarchical level as both the positive premise and the conclusion lowers argument strength.

Bayesian approaches

In the Bayesian approach to inductive reasoning, it is assumed that people make optimal inferences based on prior hypotheses about the distribution of the novel feature and the evidence provided through the premises (e.g., Heit, 1998; Tenenbaum & Griffiths, 2001). Every hypothesis about a novel feature can be formulated as the extension of the feature, i.e., which categories have it and which categories do not. The prior probability of such a hypothesis reflects the prior belief that the corresponding feature extension is correct, relative to other hypotheses. As evidence is observed (the premise), the probability of the hypotheses is updated following Bayes' rule, and the probability, that a specific category has the property, is updated accordingly.

A Bayesian inference mechanism as such does not exclude that negative evidence raises the strength of an argument, given a right set of prior probabilities for the relevant hypotheses. To raise argument strength with negative evidence, the priors of the hypotheses require a strong a priori clustering of the positive premise and the conclusion category. In other words, hypotheses that the feature extends only to the positive premise and conclusion categories and not to the negative premise should be likely a priori (i.e., before the premises are considered). Moreover, hypotheses that do not endorse this clustering should receive a low prior, for instance, hypotheses that all categories have the property or that only the positive premise category has it, or any combination of categories from the positive and negative set. Given these priors, the posterior

probability that the conclusion category has the property can increase when considering the negative premise.

Another way of looking at it is to consider an exhaustive hypothesis space. For instance, for the 'elicitation of alpha waves among music' one might have a set of hypotheses consisting of "noise in general elicits alpha waves", "only music does", "only classical music does" or "it only applies to Shostakovich". Negative evidence may help in reducing the number of hypotheses by explicitly contradicting some of them (e.g., not all music elicits alpha waves). Furthermore evidence from concept learning suggests that negative evidence even constrains the generation of hypotheses already at the outset of learning (Houtz, Moore, & Davis, 1973). Assuming a probability distribution over these hypotheses would imply an increase in probability of any of these hypotheses when excluding another hypothesis. In the example above, introducing negative evidence that excludes the noise and music in general hypotheses would hence lead to an increase in likelihood for the remaining two hypotheses. Whether this increase is large enough to replicate our empirical findings depends on the prior probabilities of each of the hypotheses.

The question then becomes how to arrive at the right prior probabilities. The priors can be considered an implementation of prior knowledge that people have regarding the domain and feature that form the topic of the argument. Heit (1998) proposes that the prior probability for each hypothesis depends on the number of familiar properties that can be retrieved from memory, and have the same extension as the hypothesis proposes: the extension of a novel feature is likely if its distribution resembles that of many already-known properties. It is, however, not immediately obvious whether a process of sampling properties can result in priors that reflect the strong clustering of categories necessary to raise the argument strength with negative evidence.

Another fruitful approach to the question about priors has been provided by the structured statistical models approach (Kemp et al., 2009; Tenenbaum, Kemp & Shafto, 2007). They

propose that the priors for the Bayesian inference derive from a stochastic process that operates on a knowledge structure. The knowledge structure (e.g., a causal food web or a taxonomic tree) captures the structure in the world that is informative for a certain argument, for instance, the taxonomic relations for inferences regarding properties in animals. If we assume a similarity representation (a tree representation or a spatial representation) and a diffusion process that distributes a feature smoothly over the structure (Kemp & Tenenbaum, 2009), as seems appropriate for the present task, the model predictions follow the Monotonicity Principle and hence cannot explain our present findings.[6] This is not to say that no combination of knowledge structure and stochastic process exists that is able to supply appropriate priors for raising argument strength with negative evidence: The structured statistical models approach is a framework that allows many specific instantiations, however speculating which of these are able to account for the present findings falls outside the scope of this paper.

Relevance theory

An approach that is not formalized but may provide a process acount to explain the present finding is the relevance theory of induction (Medin, Coley, Storms, & Hayes, 2003). The basic idea is that distinctive properties of the premise categories highlight 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. Negative evidence may work at this process of reinforcement or undermining. 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 AC/DC is 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 suggests that both the level of effort necessary to process an input and the effect that such an input has, affect the relevance of the 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.

This proposal is clearly only a rough sketch of a possible mechanism by which negative evidence may increase argument strength. As pointed out by Medin et al. (2003), the concepts of effect and effort are notoriously vague, hence making a formalization of the relevance approach a difficult task. However, the effect presented here with its proposed mechanisms within the framework of relevance, offers another way to test the relevance account of induction. Future studies may want to test whether the introduction of negative evidence can indeed highlight a dimension for induction that would otherwise not have been considered. Likewise it is an empirical question whether negative evidence can lower the effort necessary to identify a relevant basis for induction.

Conclusion

The present paper provides empirical evidence for the idea that negative evidence can increase argument strength. These findings constitute a new phenomenon of category-based property induction that models of induction need to be able to accommodate.

In their current form there are only two models of induction that can explicitly incorporate negative evidence (Blok et al., 2007; Kemp & Tenenbaum, 2009). Employing minimal assumptions, we have offered a brief overview of the main models and discussed whether or how they could incorporate our findings. Those models solely based on the underlying similarity relations between premises and conclusion seem to have a hard time accounting for the increase in argument strength. Bayesian models are better able to accommodate the phenomenon, however they require a specific distribution of prior probabilities across hypotheses in order to explain the effect. Thus the onus is on them to provide a reasonable mechanism that would result in these priors. The relevance theory provides an intuitive process account of how the present effect may come about. However, relevance theory is only a framework account that is not formalized. The relatively vague concepts of effort and effect afford a larger degree of flexibility and can hence easily accommodate a range of phenomena.

Phenomena like the one presented here provide the opportunity to challenge underlying assumptions of models of induction. Deriving mechanisms to accommodate these findings leads to new predictions that in turn provide tests of the models. The common goal of all models of induction is to provide a sensible way to define the relevance of evidence for a conclusion. The violation of the Monotonicity Principle for negative evidence constitutes a clear constraint on models of induction and again highlights the importance of getting a clearer grasp of what determines the relevance of evidence—be it positive or negative.

*References

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

Feeney, A. (2007). How many processes underlie category-based induction? Effects of conclusion specificity and cognitive ability. *Memory & Cognition, 35,* 1830-1839.

Grice, H. P. (1975/1989). Logic and conversation. In H. P. Grice (Ed.), *Studies in the way of words* (pp. 22-40). Cambridge, MA: Harvard University Press.

Hampton, J.A. & Cannon, I. (2004). Category-based induction: An effect of conclusion typicality. *Memory & Cognition, 32*, 235-243.

Hayes, B., Heit, E., & Swendsen, H. (2010). Inductive reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 278-292.

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.

Heit, E. Hahn, U. & Feeney, A. (2004) Defending diversity. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff, (Eds.), *Categorization inside and outside of the lab: Festschrift in honor of Douglas L. Medin*. Washington, DC: APA Press.

Hempel, C. G. (1945) Studies in the Logic of Confirmation I. *Mind, 54,* 1–96.

Heussen, D. & Hampton, J. A. (2011). Induction with mixed evidence: The role of typicality. *Under review.*

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.

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 generalization. *Thinking & Reasoning, 13,* 394-425.

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

McDonald, J., Samuels, M., & Rispoli, J. (1996). A hypothesis assessment model of categorical argument strength. *Cognition, 59,* 199-217.

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,* 185–200.

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

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317-1323.

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,* 67-96.

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

Tenenbaum, J. B. & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences, 24,* 629.

Tenenbaum, J. B., Kemp, C. & Shafto, P. (2007). Theory-based Bayesian models of inductive reasoning. In Feeney, A. & Heit, E. (Eds.), *Induction*. Cambridge, U.K.: Cambridge University Press.

Waxman, S.R., Lynch, E.B., Casey, K.L., & Baer, L. (1997). Setters and samoyeds: The emergence of subordinate level categories as a basis for inductive inference. *Developmental Psychology*, *33*, 1074–1090.

Appendix: Items used in Experiments 1 & 3

	Positive Premise	Negative Premise	Conclusion	Exp. 1 N=32	Exp. 3 N=14	Exp. 3 Positive	Exp. 3 Mixed	Grouping task
Target Items	French wine,	¬Australian wine	→ Italian wine	0.16	0.14	6.43	6.36	0.43
	Rabbit,	¬Hedgehog	→ Squirrel	0.25	0.29	7.00	6.79	0.43
	BMW,	¬Fiat	→ Mercedes	0.28	0.29	6.71	6.64	0.93
	Duck,	¬Sparrow	→ Swan	0.38	0.36	7.07	7.79	0.57
	Flute,	¬Guitar	→ Trumpet	0.47	0.29	5.36	6.07	0.50
	Rubens,	¬Dali	→ Van Eyck	0.50	0.21	4.79	4.50	0.93
	Actor,	¬Librarian	→ Politician	0.50	0.43	4.86	5.86	0.79
	Freight ship,	¬Hovercraft	→ Cruise ship	0.50	0.57	7.00	8.07	0.93
	Atlantic,	¬Lake Balaton	→ Mediterranean	0.50	0.29	5.71	6.36	0.86
	Window glass,	¬Bottle glass	→ Car glass	0.53	0.14	5.43	5.57	0.86
	Polar bear,	¬Bison	→ Penguins	0.53	0.57	7.07	8.29	1.00
	Mozart,	¬AC/DC	→ Bach	0.59	0.36	6.43	7.07	1.00
	Moth,	¬Spider	→ Fly	0.63	0.21	5.64	6.07	0.93
	LCD TV,	¬Old TV set	→ Plasma screen	0.69	0.57	6.64	7.86	0.43
	Strawberry,	¬Banana	→ Raspberry	0.75	0.43	6.86	7.36	0.93
Control Items	Guitarist,	¬Basguitarist	→ Violinist	0.06	0.00	7.71	4.50	0.00
	Laptop PC,	¬Palmtop PC	→ Calculator	0.06	0.14	5.50	3.57	0.14
	Air gun,	¬Sniper rifle	→ Hunting rifle	0.13	0.21	6.43	6.07	0.79
	Ant,	¬Termite	→ Bee	0.13	0.00	4.64	3.43	0.50
	Potato,	¬Beet	→ Carrot	0.13	0.07	5.50	4.14	0.43
	Horse,	¬Cow	→ Goat	0.13	0.00	5.36	2.71	0.36
	Papaya,	¬Star fruit	→ Mango	0.16	0.00	5.93	5.21	0.43
	F16,	¬Concorde	→ Boeing	0.19	0.07	6.21	4.43	0.29
	Oak,	¬Willow	→ Beech	0.19	0.07	5.07	4.07	0.57
	Swordfish,	¬Ray	→ Tuna	0.19	0.00	5.00	4.21	0.64
	Stork,	¬Crow	→ Goose	0.22	0.07	6.00	4.07	0.79
	Picasso,	¬Magritte	→ Warhol	0.28	0.07	5.43	3.93	0.71
	Lion,	¬Coyote	→ Crocodile	0.28	0.00	5.00	3.71	0.71
	Tripel beer,	¬Dubbel beer	→ Duvel beer	0.31	0.00	6.21	4.71	0.64
	Snake,	¬Wasp	→ Scorpion	0.31	0.21	5.93	5.50	0.50

Note. Columns headed Exp. 1 & 3 contain the proportion of people indicating a preference for arguments containing negative evidence. The columns headed positive and mixed show the ratings of single positive and mixed premise arguments on a scale from zero to ten. The last column shows the proportion of participants (N = 14) who selected the positive premise and the conclusion in a grouping task.

Appendix: Proportion of people preferring arguments containing negative evidence. Item scores for each condition in Experiment 2.

	Positive Premise	Negative Premise	Conclusion	Target N=41	Control N=41	Irrelevant N=39
Target Items	Rubens,	¬Dali	→ Van Eyck	0.37	0.22	0.23
	Actors,	¬Librarian	→ Politicians	0.39	0.07	0.21
	Freight ship,	¬Hovercraft	→ Cruise ship	0.41	0.17	0.31
	Atlantic,	¬Lake Balaton	→ Mediterranean	0.27	0.05	0.15
	Window glass,	¬Bottle glass	→ Car glass	0.34	0.34	0.28
	Polar bear,	¬Bison	→ Penguin	0.54	0.27	0.33
	Mozart,	¬AC/DC	→ Bach	0.66	0.10	0.44
	Moth,	¬Spider	→ Fly	0.37	0.12	0.33
	LCD TV,	¬Old TV set	→ Plasma screen	0.66	0.37	0.31
	Strawberry,	¬Banana	→ Raspberry	0.34	0.02	0.36

Note. Columns contain the proportion of people indicating a preference for arguments containing negative evidence in each condition. Sample size for each condition is given at the top.

Author Note

This work was supported by an F+ Fellowship awarded to the first author by the Universit of Leuven. Enquiries concerning this paper should be addressed to Daniel Heussen or Wouter Voorspoels, Department of Psychology, K.U. Leuven, Tiensestraat 102, 3000 Leuven, Belgium, daniel.heussen@psy.kuleuven.be; wouter.voorspoels@psy.kuleuven.be.

Footnote

1. In fact, rather than using explicit negation, negative evidence is sometimes implemented as having another type of property (e.g., Kalish & Lawson, 2007).
2. In order to test the authors' intuition participants performed an additional grouping task in Experiment 3. Participants were asked to circle the two exemplars out of the triplet present in the argument that belong together. In the appendix, the last column of the table shows the proportion of participant indicating the first premise and the conclusion as the grouped pair. For the target condition, 10 out of 15 items had the majority of participants (around .8 and above) select the first premise and conclusion as the grouped pair (overall average .77 for the target condition). In contrast, across items the average proportion of people selecting the first premise and conclusion for the control condition was .5.
3. The No-change responses were omitted from the analyses to avoid violations of the independence assumption of ANOVA.
4. Although we restrict ourselves in referring to the effect on people's preferences, the effect held across both people and items as indicated by the min*F*' analyses.
5. Furthermore the similarity component uses the maximum similarity of all premises to the conclusion. In mixed premise arguments, the sign of the similarity component would hence depend on whether the positive or the negative premise is more similar to the conclusion leading to unnatural patterns of data (Heussen, Voorspoels & Storms, 2010).
6. Also, contrary to the non-monotonicity effects when adding positive evidence (Kemp et al, 2009; Tenenbaum & Griffiths, 2001), varying sampling assumptions (strong or weak sampling, the size principle) does not explain the present findings.

Figure Captions

Figure 1. Average proportion of responses selecting the mixed premise argument for target and control items. Error bars constitute 95% confidence intervals calculated on the participant data.

Figure 2. Average proportion of responses selecting the mixed premise argument for target, control and irrelevant items. Error bars constitute 95% confidence intervals calculated on the participant data.

Figure 3. Average proportion of responses indicating a change in argument strength from the single positive (first judgment) to mixed premise arguments (second judgment) for both control and target items. Error bars constitute 95% confidence intervals calculated on the participant data.

Figure 4. Average argument strength of single positive (first judgment) and mixed premise arguments (second judgment). Error bars constitute 95% confidence intervals calculated on the participant data.

Figure 1

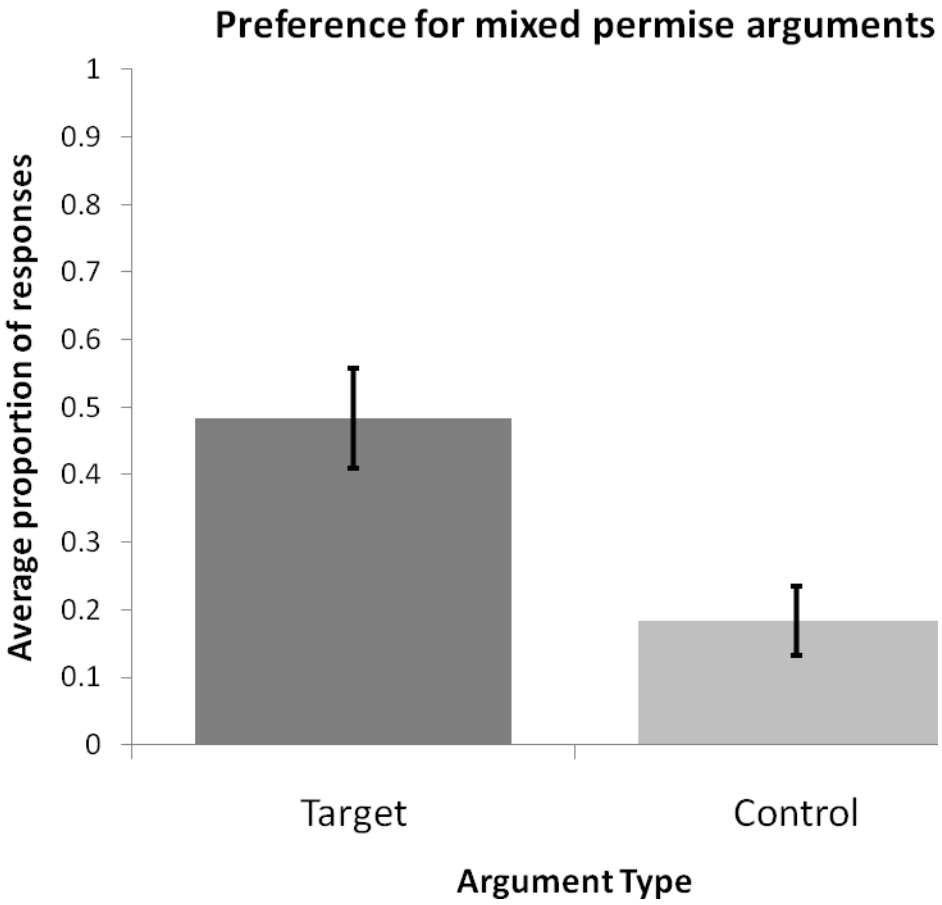


Figure 2

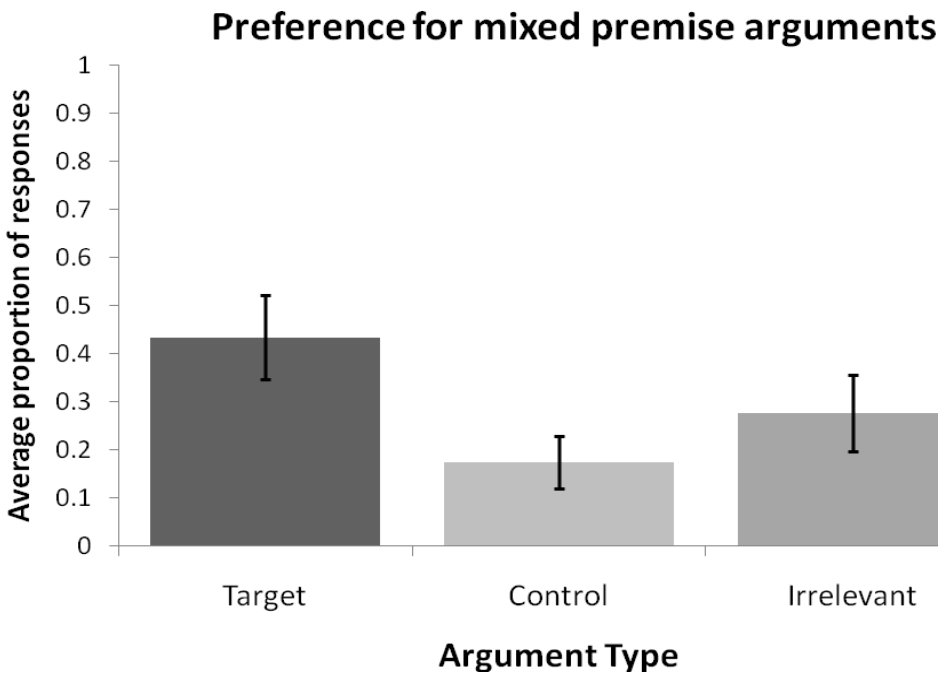


Figure 3

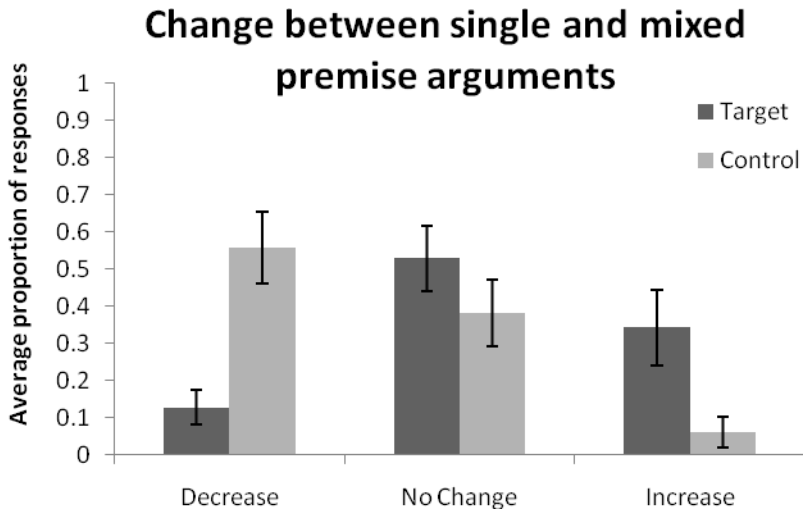
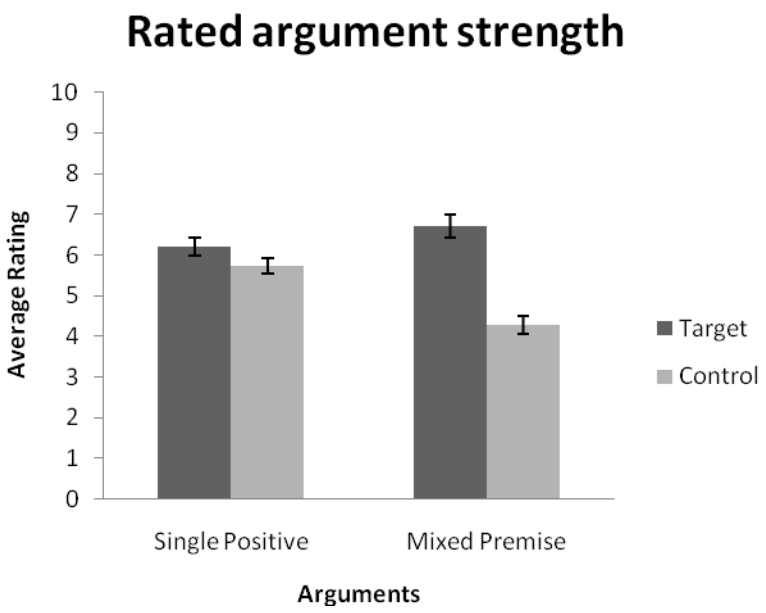


Figure 4


Rated argument strength
Average Rating
10
9
8
7
6
5
4
3
2
1
0
Single Positive
Mixed Premise
Arguments
Target
Control