Ransom, K., Perfors, A. and Navarro, D. J. (2016). Leaping to conclusions: Why premise relevance affects argument strength. Cognitive Science, 40, 1775-1796 https://doi.org/10.1111/cogs.12308

Leaping to conclusions: Why premise relevance affects argument strength

Keith Ransom
Amy Perfors
Danielle J. Navarro
School of Psychology
University of Adelaide

Abstract

Everyday reasoning requires more evidence than raw data alone can provide. We explore the idea that people can go beyond this data by reasoning about how the data was sampled. This idea is investigated through an examination of *premise non-monotonicity*, in which adding premises to a category-based argument weakens rather than strengthens it. Relevance theories explain this phenomenon in terms of people's sensitivity to the relationships amongst premise items. We show that a Bayesian model of category-based induction taking premise sampling assumptions and category similarity into account complements such theories and yields two important predictions: first, that sensitivity to premise relationships can be violated by inducing a weak sampling assumption; and second, that premise monotonicity should be restored as a result. We test these predictions with an experiment that manipulates people's assumptions in this regard, showing that people draw qualitatively different conclusions in each case.

Keywords: Bayesian modelling; category-based induction; non-monotonicity; relevance theory; sampling assumptions.

Correspondence concerning this article should be addressed to Keith Ransom, School of Psychology, University of Adelaide SA 5005, Australia (keith.ransom@adelaide.edu.au). DJN received salary supported from ARC grant FT110100431 and AP from ARC grant DE120102378. Research costs were funded through ARC grant DP110104949.

Introduction

Whereas formal deductive reasoning provides a solid bridge from premise to conclusion, everyday reasoning requires an inferential leap. But what assumptions support such a leap when raw data alone cannot? This question is relevant to the understanding of category-based induction, an important and representative form of inductive reasoning. In a typical category-based induction task, people are presented with a conclusion supported by one or more premise statements and asked to rate the strength of the inductive argument as a whole. Similarity-based models, which assume that argument strength is assessed on the basis of similarity between premise and conclusion categories, have successfully accounted for many aspects of people's performance in such tasks (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Sloman, 1993). Yet there are other characteristics of people's reasoning in this regard that are not adequately predicted on the basis of similarity. These characteristics have been explained as emerging from people's sensitivity to the relevance of different premises and the relationships amongst them (Medin, Coley, Storms, & Hayes, 2003).

In this paper we explain why and when premise relevance should matter. We argue that people's reasoning is sensitive to premise relationships because they consider the generative process behind the data they observe. If people made no such considerations, and instead assumed that all data consistent with the truth were equally likely to be observed (a so-called *weak sampling* assumption), then a perceived relationship amongst premise items should have no effect on argument strength. We demonstrate this by manipulating people's assumptions about premise selection, and observing that people draw qualitatively different conclusions as a result. Thus, we reproduce an effect of premise relevance on argument strength demonstrated by Medin et al. (2003) in a scenario where relevance should matter, and fail to observe the effect where it should not. Furthermore, we argue that the notion of *cognitive effect*, central to relevance theory explanations of induction, is neatly captured by a Bayesian theory of category-based induction that naturally incorporates different assumptions about premise sampling, along with the role of category similarity. Our results offer important corroborating evidence for the relevance theory of induction.

We first describe the category-based induction task, with a focus on arguments in which additional premises lead to weaker rather than stronger conclusions (known as *premise non-monotonicity*). We then describe a Bayesian analysis of this task which predicts that whether or not people exhibit premise non-monotonicity depends critically on how they assume the premises were generated in the first place. Finally, we present an experiment in which we manipulate these assumptions. As predicted by our model, people's reasoning differs qualitatively as a function of how they think the premises were sampled.

Premise monotonicity and non-monotonicity

In a typical category-based induction task participants are asked to rate the strength of inductive arguments like the following:

premise	EAGLES have more than one fovea per eye.
conclusion	HAWKS have more than one fovea per eye.

Here we use the notation EAGLES → HAWKS to indicate that this problem asks people to generalize a property from EAGLES to HAWKS.[1] Given that EAGLES and HAWKS are similar, participants might rate this as a moderately strong argument. Adding premises to an argument typically strengthens it, an effect referred to as *premise monotonicity* (Osherson et al., 1990). For instance, the argument {EAGLES, FALCON} → HAWKS appears stronger than EAGLES → HAWKS. The additional premise provides evidence that the property of *multiple foveae* should be extended to all birds of prey, and is not a property of EAGLES alone.

However, systematic violations of premise monotonicity have been observed. For example, Medin et al. (2003) found that people were less willing to endorse the generalization {GRIZZLY BEARS, BROWN BEARS, POLAR BEARS} → BUFFALO than GRIZZLY BEARS → BUFFALO, despite the former having more premises. This *non-monotonicity* effect appears to arise because the multiple premise argument provides strong evidence that the property should be extended to bears only, and so weakens the plausibility that buffaloes share the property. This insight is captured in the relevance theory of induction, which suggests that adding premise categories should weaken an argument if the added categories reinforce a property shared by all of the premise categories but not the conclusion (Medin et al., 2003).

This seems sensible, but *why* is it so? If nothing can be assumed about the way premises are sampled, then there is no reason to expect a more relevant premise to be advanced in argument over a less relevant one; the notion that a perceived relationship between premise items represents the appropriate basis for induction, gains no special credence simply by virtue of being put forward. But in the real world arguments are rarely (if ever) constructed from randomly sampled facts. It makes sense for people to assume that arguments are constructed by sampling relevant facts to support conclusions and achieve communication goals. Wilson and Sperber's account of *relevance theory* (Wilson & Sperber, 2004) and Grice's *co-operative principle* (Grice, 1989), upon which their theory is based, each offer explanations for why utterances raise an expectation of relevance on the part of the listener. For Grice, the raised expectation comes about because people, for the most part, follow communicative conventions that encourage relevance. But such a heightened expectation should serve only to sharpen the ability to discriminate inputs on the basis of relevance. A reasonable variation in relevance should exist in the first place.

[1] More precisely, we might denote this EAGLES $\xrightarrow{\textit{mult. foveas}}$ HAWKS in order to emphasize the property being extended in the argument. For the most part this detail is not needed for our paper.

Wilson and Sperber go further than Grice, arguing that neither a communicative convention nor a communicative context are strictly necessary for an enhanced perception of relevance. A tendency to maximize relevance, they contend, is a fundamental feature of our cognitive systems, arising from the need to make the most efficient use of processing resources. To give an example, there are a number of theoretical results showing that positive evidence has stronger evidentiary value than negative evidence under plausible assumptions[2] about the environment (e.g., Klayman & Ha, 1987; Navarro & Perfors, 2011). Given this, maximizing relevance should lead people to prefer to give and to receive positive evidence, and will therefore treat positive premises (of the form "item x has property p") as more relevant than negative ones.

If people assume premises are sampled based on relevance then any property shared by the premises will gain plausibility as the correct basis for induction and stronger inferences to that effect should result. For example, if I want to convey the range of animals that share a particular property, and I want to be as relevant as possible, then I should select additional examples that best capture the appropriate range. Returning to the bears example, had I wanted to convey the message that many species had some property, not just bears, you might reasonably have expected me to mention a different kind of animal. So my choosing further examples of bears when extending my argument provides evidence that only bears have the property. Qualitatively, this reasoning explains why people exhibit premise non-monotonicity in this situation. This intuition can be reinforced quantitatively by the mathematics of Bayesian probability theory, as we explain in the next section.

A model for reasoning in category-based induction

Consider a standard Bayesian approach to category-based induction tasks (Heit, 1998; Sanjana & Tenenbaum, 2003). Suppose the learner is given a one premise argument of the form $x \xrightarrow{p} y$. Let h denote one possible hypothesis about how far property p should be extended, and $P(h)$ denote the reasoner's prior bias to think that h describes the true extension of property p. Having observed that item x possesses property p, the posterior degree of belief in h is given by Bayes' rule:

$$P(h \mid x) = \frac{P(x \mid h)P(h)}{\sum_{h'} P(x \mid h')P(h')}. \tag{1}$$

[2]The critical assumption is that we live in a world in which most items do not have most properties. This seems intuitive (e.g., FOXES are *furry*, but FISH, FEARS and FOOTPRINTS are not), but some care is needed in substantiating the point. From a logical standpoint any "sparse" property (possessed by a minority of entities) is mirrored by a "non-sparse" complement. However, they need not be equally salient nor equally useful when describing the world: people are more likely to think that *furry* (sparse) as a meaningful property than *non-furry* (non-sparse). Indeed, what Navarro and Perfors (2011) show is that in any world where entities are not completely homogeneous, the categories and properties that intelligent agents attend to should display this sparsity bias.

Here, $P(x \mid h)$ is the likelihood, which specifies the probability that the argument would have used x as a premise if h were the true extension of property p. The sum in the denominator is taken over all hypotheses that the reasoner might consider regarding the extension of property p. When an argument contains multiple premise items $x_1, \ldots, x_k$, the likelihood is given by the product of each of the individual probabilities, $\prod_{i=1}^{k} P(x_i \mid h)$. In order to evaluate the claim that item y also possesses property p, a Bayesian reasoner sums the posterior probabilities of all hypotheses that are consistent with the claim. Thus, the argument strength is given by:

$$P(y \mid x) = \sum_{h: y \in h} P(h \mid x). \tag{2}$$

This model has two components, the prior $P(h)$ and the likelihood $P(x|h)$. In our application of the model, the prior reflects the similarity amongst premise categories. As described in Appendix A, we use empirical similarity data to set $P(h)$ and simulations to check that the qualitatively important effects are not overly sensitive to the particular data collected.

The likelihood is critical to an understanding of when and why premise relevance matters: it naturally captures different assumptions people may make about how the premises were generated. For instance, a naive reasoner might assume that the premise items for an argument are selected at random from the set of true facts about the property p. This is called *weak sampling*. Since the item x is chosen randomly, weak sampling allows premises to present negative evidence (i.e.,"item x does not have property p"). For a premise presenting evidence that item x has property p, the weak sampling likelihood function is:

$$P(x \mid h) \propto \begin{cases} 1 & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In essence, when presented with item x, a learner assuming weak sampling falsifies all hypotheses inconsistent with the premise but does not alter their beliefs in any other respect – the fact that x was chosen over other items has no additional relevance to the reasoning problem. As a result, such a learner will be less likely to demonstrate premise non-monotonicity. If premises are generated randomly, seeing BLACK BEARS in addition to GRIZZLY BEARS does not act as a "hint" that only bears have the property in question. Rather, because there are almost no hypotheses that could be falsified by the additional BLACK BEARS premise that were not already falsified by the GRIZZLY BEARS premise, the additional information is largely irrelevant.

The simplicity of the weak sampling model and its connection to falsification is appealing. However, as we have seen, it provides a poor description of how inductive arguments are constructed in everyday reasoning. If a learner expects an argument to be constructed using positive examples, then a weak sampling assumption is no longer tenable. A simple alternative is *strong sampling* (Tenenbaum & Griffiths, 2001; Sanjana & Tenenbaum, 2003), in which a premise item is selected *only* from those

exemplars that possess property p. As noted earlier, this restriction makes sense if people expect to receive relevant evidence. This gives the likelihood function

$$P(x \mid h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $|h|$ denotes the *size* of hypothesis h. In this context, the size is calculated by counting the number of items that possess property p assuming hypothesis h is true.

Under strong sampling, the item presented has relevance beyond falsification. That is, a premise provides more evidence for a small hypothesis than it does for a larger one (Tenenbaum & Griffiths, 2001). A learner who sees multiple premise items consistent with one small hypothesis will come to prefer that hypothesis over other, broader hypotheses, even when the broader hypothesis happened to be originally preferred. As noted in previous research (Fernbach, 2006; Kemp & Tenenbaum, 2009; Voorspoels, Van Meel, & Storms, 2013), this phenomenon provides a potential explanation for why people sometimes exhibit premise monotonicity and at other times non-monotonicity.

Compare the one premise argument CHIMPANZEES → GORILLAS to the two premise argument {CHIMPANZEES, ORANGUTAN} → GORILLAS. Both premises are consistent with a small hypothesis (i.e., that all primates have that property). Because gorillas are also primates the additional evidence provided by the ORANGUTAN premise acts to strengthen the argument: premise monotonicity is satisfied. In contrast, compare the one premise argument GRIZZLY BEARS → LION to the two premise argument {GRIZZLY BEARS, BLACK BEARS} → LION. In the one premise variant, the reasoner might reasonably believe that the property extends to all mammals or all predators, and so there is at least some chance that LIONS possess the property. However, when BLACK BEARS is added to the list of premise items, the reasoner has strong evidence in favor of a small hypothesis, namely that the property is common only to bears. This produces a non-monotonicity effect, since an additional positive observation acts to weaken the conclusion.

Importantly, this explanation relies on the assumption of strong sampling. It is this assumption that gives a premise item relevance over and above its use for falsification. In the bears example above, the effect occurs because a second bear premise provides strong evidence for the (smaller) "bears" hypothesis relative to the (larger) "all predators" and "all mammals" hypotheses, even though all three are consistent with both premises. Under a weak sampling assumption, premise items have no relevance beyond falsification, and this shift does not occur.

If strong sampling represents an assumption that premise selection is biased towards relevant items[3] and weak sampling represents the assumption that is is not,

[3]The strong sampling model is not intended to capture all the complexity of selecting items for relevance. For instance, richer pragmatic assumptions can be captured using pedagogical sampling models (Shafto, Goodman, & Griffiths, 2014). This complication is not necessary in the current context but some implications are addressed in more detail in the discussion.

then it is reasonable to consider that the bias to expect relevant premises might vary not just in kind, but also in degree. A *mixed sampling* model can be used to capture this situation in a straightforward way (Navarro, Dry, & Lee, 2012). Under mixed sampling, the likelihood function becomes:

$$P(x \mid h) = \begin{cases} \theta \frac{1}{|h|} + (1-\theta)\frac{1}{|\mathscr{X}|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $|\mathscr{X}|$ represents the number of possible premise items, and θ represents the probability that the premise item x was strongly sampled. When $\theta = 0$ the model is equivalent to weak sampling and has no bias towards positive evidence. In contrast, when $\theta = 1$ the bias is so extreme that the learner believes it is impossible to receive negative evidence, and the mixed sampling model becomes equivalent to strong sampling.

The notion that people are sensitive to how the premises were generated represents an intriguing and testable prediction. If reasoners have an expectation of premise relevance and thus expect premises to be biased towards positive evidence, they should show premise non-monotonicity for the bears example. If, on the other hand, they assume that premise items have been selected at random (i.e., weakly sampled), then premise monotonicity should be exhibited. Note that this prediction stands in contrast to the predictions of similarity based models (Osherson et al., 1990; Sloman, 1993) neither of which incorporate any sensitivity to the mechanism by which the premises are generated. To that end, we present experimental evidence that premise monotonicity can be systematically manipulated by changing the assumptions people make about the origins of the data. Not only do we see qualitative reversals from monotonic to non-monotonic reasoning consistent with a change from weak to strong sampling, we also find that the transition occurs in a graded fashion consistent with the smoothly varying bias parameter in the mixed sampling model.

Experiment

Method

Participants. 590 adults were recruited via Amazon Mechanical Turk, and were each paid $0.50 (USD) for the 5–10 minutes participation. 52 were excluded due to browser incompatibility, and the remaining 538 were aged 18 to 69 years (median age 28, 65% male). 500 participants were in the United States, with 38 located elsewhere.

Procedure. A cover story informed people that they would be making judgments concerning well established facts about the properties of animals. Each trial began by presenting a fact about one animal and then asking about a second. For example, they might first be told that grizzly bears produce the hormone TH-L2, and then asked whether lions also produce TH-L2. Responses were collected using a slider bar that allowed people to produce answers ranging from “100% false” to “100% true”,

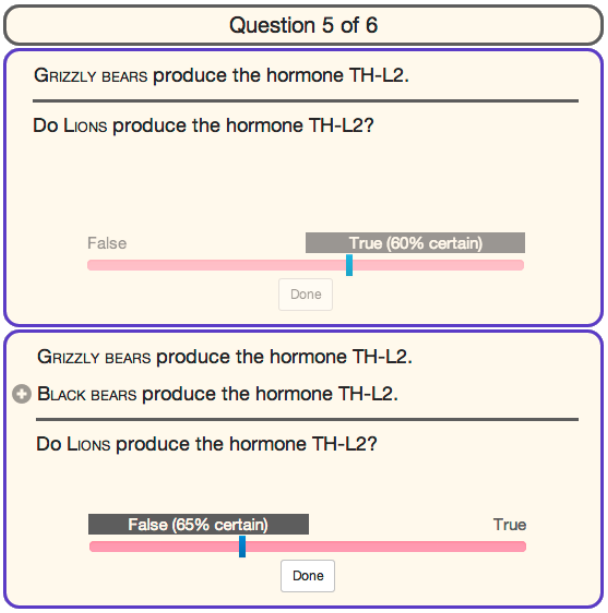


Figure 1. An illustration of the on-screen presentation of a trial shown at the point where the second premise has been revealed. The one premise argument is displayed on the upper portion of the display, while the two premise form is on the lower portion. The rectangular "slider" (disabled on the upper portion, enabled in the lower) allows participants to respond "True" or "False" and indicate the level of certainty in their response.

as shown in Figure 1. They were then told about a second animal, and asked to revise their original judgment by moving a different slider. The dependent measure for each trial is the difference between these two judgments. If the endorsement of the conclusion is stronger on the second occasion, premise monotonicity is satisfied. If the difference is negative, non-monotonicity is observed.

Conditions. Participants were randomly assigned to one of the four conditions, each involving a different combination of cover story and filler trials. The cover story informed people about how the second fact in each trial was generated, while the filler trials were designed to be consistent with either a strong or weak sampling assumption. In the BOTH RELEVANT condition, participants were told that the extra facts were provided by past players of the game who were trying to select a helpful example of an animal with the property in question. The story and the filler trials were designed to promote the idea that facts were chosen on the basis of relevance, similar to strong sampling. In the BOTH RANDOM condition people were told that they would select a card from a deck displayed face down on-screen. This card would disclose whether or not a particular animal had the property in question. In contrast to the BOTH RELEVANT condition, the story and filler items were designed to support the assumption of weak sampling by encouraging the belief that facts were

			Additional example	
Trial	Property to be generalized	First generalization	Relevant	Random
Filler 1	have more than one fovea per eye	eagles → doves	+hawks	−tortoises
Filler 2	have mammary glands	elephants → deers	+cows	+anteaters
Target 1	have a bite force greater than 500 BFU	tigers → ferrets	+lions	+lions
Filler 3	give birth to underdeveloped young	kangaroos → wombats	+koalas	−flamingos
Target 2	produce the hormone TH-L2	grizzly bears → lions	+black bears	+black bears
Control	require cystocholamine for brain function	orangutans → gorillas	+chimpanzees	+chimpanzees

Table 1: The property to be generalized, the first generalization, and additional example used in the Both Relevant/Relevant Fillers conditions, and in the Both Random/Random Fillers condition. Trials are shown are shown in the order presented in the experiment. All conditions have the same arguments in the key trials (**Target 1**, **Target 2**, and **Control**), differing only in cover story and supporting filler trials. The second generalization that people were required to make is formed by combining the first generalization with the additional example. For example, the second generalization for **Target 1** becomes {tigers, +lions} → ferrets. The "−" symbol is used to indicate that the statement should be negated: e.g., "tortoises *don't have* more than one fovea per eye."

being sampled at random. To allow us to investigate whether the premises alone had an effect on sampling assumptions we ran two further experimental conditions. The Relevant Fillers condition employed a neutral cover story giving no information about how the premises were selected, and used the same filler items as the Both Relevant condition. Likewise, the Random Fillers condition employed a neutral cover story, but used the same filler items as the Both Random condition.

Stimuli. All participants were presented with six trials in a fixed order,[4] as shown in Table 1. Three of these were especially key and appeared in all conditions. There were two target arguments structured so that they should elicit non-monotonic responding under a strong sampling assumption (**Target 1**: {tigers, lions} → ferrets; **Target 2**: {grizzly bears, black bears} → lions). There was also a **Control** argument designed to elicit monotonic reasoning under any mixture of weak or strong sampling ({chimpanzee, orangutan} → gorilla).Finally, each person saw three **Filler** trials, designed to reinforce a particular sampling assumption. Consistent with strong sampling, the filler trials in the Both Relevant and Relevant Fillers conditions consisted solely of positive examples. In contrast, the filler trials in the Both Random and Random Fillers conditions included negative examples as well, and appeared much more random.

[4]Randomization of trial order would not have made sense in this context. Because the filler items were an important part of the experimental manipulation, it was critical that at least some of these precede the target items; because we did not want the design to be too obvious, we also wanted to include at least one filler in between the two targets.

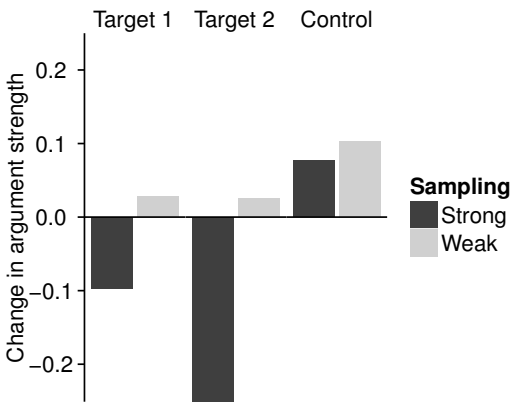


Figure 2. Model predictions for the change in argument strength when an additional premise is introduced (i.e., $P(y|x_1, x_2) - P(y|x_1)$). A positive change indicates premise monotonicity, a negative change, non-monotonicity. In the **Control** argument, monotonicity is predicted regardless of sampling assumption. For both **Target** arguments, a reversal is predicted: premise non-monotonicity is expected only under an assumption of strong sampling. (The difference in the magnitude of the predictions between the two **Target** conditions emerges due to the structure of people's real-world knowledge about the domain as reflected in the prior, and is incidental to our main point.)

Results

Model predictions

The Bayesian model of strong and weak sampling described in Equations (1) to (4) was used to quantitatively predict how a reasoner holding either assumption would reason about the two **Target** arguments and the **Control** argument. In order to extract these predictions, it was necessary to specify a hypothesis space $\mathcal{H}$ and a prior distribution $P(\mathcal{H})$. The hypothesis space simply consisted of all possible sets of the 14 animals common to the two experimental conditions. In order to estimate the prior, we collected similarity ratings for all pairs of the 14 animals. The estimation procedure was an adaptation of the additive clustering technique (Shepard & Arabie, 1979; see also Lee, 2002, Navarro & Griffiths, 2008) and is discussed in more detail in Appendix A.

Figure 2 shows the resulting model behavior. As predicted previously, when weak sampling is assumed the model indicates premise monotonicity for both **Target** and **Control** trials. Conversely, under strong sampling it predicts non-monotonicity for **Target** trials and monotonicity for **Control** trials. Importantly, while the precise numerical prediction shown in Figure 2 depends on the way in which the prior was derived, the qualitative effect of sampling assumptions is robust with regard to change in details: as discussed in Appendix A, the Bayesian model predicts a shift towards

		Argument strength					
		Original		Revised		Change	
Condition	N	Mean	SE	Mean	SE	Mean	SE
Target 1							
Both Relevant	135	.283	.021	.210	.021	-.073	.013
Relevant Fillers	134	.313	.023	.259	.022	-.054	.015
Random Fillers	138	.301	.020	.275	.020	-.026	.014
Both Random	131	.277	.022	.307	.026	.031	.021
Target 2							
Both Relevant	135	.523	.015	.444	.023	-.079	.023
Relevant Fillers	134	.538	.017	.484	.022	-.054	.015
Random Fillers	138	.534	.017	.521	.020	-.012	.014
Both Random	131	.578	.018	.616	.021	.038	.013
Control							
Both Relevant	135	.773	.015	.863	.014	.090	.013
Relevant Fillers	134	.765	.015	.860	.013	.096	.009
Random Fillers	138	.759	.013	.853	.013	.093	.011
Both Random	131	.790	.016	.902	.010	.111	.013

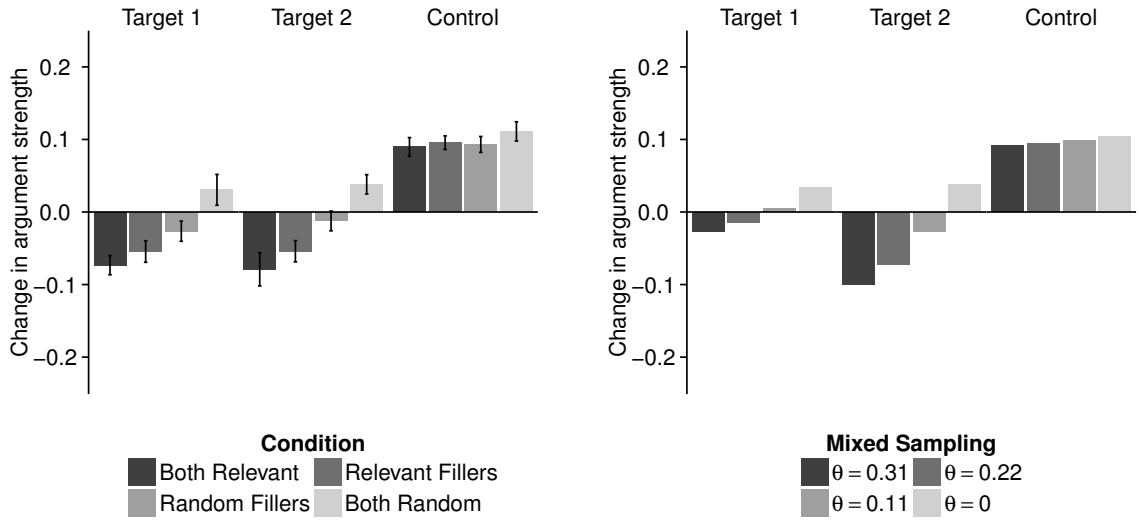
Table 2: Mean argument strength ratings (linearly scaled to the range 0 to 1) for the original judgment (after seeing the first premise only), the revised judgment (after seeing the second premise), and mean change in argument strength (the revised rating minus the original rating, linearly scaled to the range -1 to 1), summarised by condition and trial type.

non-monotonicity under strong sampling provided that the prior distribution reflects the conceptual structure of the animal domain.

Experimental results

For each trial, participants rated the strength of an argument in a one- and two-premise form. The main question of interest was whether sampling assumptions had an impact upon the way people assessed the evidentiary value of the additional premise. The dependent measure was therefore the response change between the two judgments: a positive response change reflects premise monotonicity, while a negative one reflects non-monotonicity. Table 2 presents mean argument strength ratings based on the one- and two-premise forms, as well as the mean change between judgments, by trial type and condition.

Figure 3(a) shows, as predicted, that people exhibited different response patterns depending on their sampling assumptions. For both **Target** trials, participants in the Both Random condition exhibited premise monotonicity, while those in the Both Relevant condition showed non-monotonicity. To quantify the amount of evidence for these assertions, for every condition we ran Bayesian analysis comparing



(a) Human performance (b) Model fits

Figure 3. (a) Average change in people's argument strength ratings for all four conditions, calculated by subtracting their original judgment (after seeing the first premise only) from their revised judgment (after seeing the second premise), then linearly scaled to the range -1 to 1. In keeping with the predictions, people exhibit premise non-monotonicity in the Both Relevant and Relevant Fillers conditions and only for the **Target** arguments. The results demonstrate that when a relationship amongst premise categories not shared by the conclusion is highlighted, a strong reason is needed in order for such relevance to be ignored and for non-monotonic reasoning to be inhibited. Bars show one standard error. (b) Best fitting value of θ under a mixed sampling assumption. $\theta = 0$ corresponds to a weak sampling assumption, whereas $\theta = 1$ would correspond to an assumption of pure strong sampling. Intermediate values reflect more graded assumptions. The fitted values confirm that when a cover story establishes a high or low expectation of premise relevance consistent with the premises observed, people exhibit an increased bias towards strong or weak sampling, respectively.

Condition	Bayes Factor		
	Target 1	Target 2	Control
Both Relevant	$> 1{,}000 : 1$ ($\boldsymbol{\mu < 0} : \mu = 0$)	$40 : 1$ ($\boldsymbol{\mu < 0} : \mu = 0$)	$> 1{,}000 : 1$ ($\boldsymbol{\mu > 0} : \mu = 0$)
Relevant Fillers	$98 : 1$ ($\boldsymbol{\mu < 0} : \mu = 0$)	$91 : 1$ ($\boldsymbol{\mu < 0} : \mu = 0$)	$> 1{,}000 : 1$ ($\boldsymbol{\mu > 0} : \mu = 0$)
Random Fillers	$1 : 1$ ($\mu < 0 : \mu = 0$)	$1 : 5.7$ ($\mu < 0 : \boldsymbol{\mu = 0}$)	$> 1{,}000 : 1$ ($\boldsymbol{\mu > 0} : \mu = 0$)
Both Random	$1 : 1.8$ ($\mu > 0 : \boldsymbol{\mu = 0}$)	$13 : 1$ ($\boldsymbol{\mu > 0} : \mu = 0$)	$> 1{,}000 : 1$ ($\boldsymbol{\mu > 0} : \mu = 0$)

Table 3: Bayes factors indicating the relative likelihood of a one-sided model of mean change in argument strength against the null model, by condition and trial type. The one-sided test performed in each case (given in parentheses) was chosen on the basis of the mean change in argument strength observed. $\mu < 0$, $\mu = 0$ and $\mu > 0$ correspond to the hypotheses that the true mean change in argument strength represents non-monotonic, strictly flat and monotonic responding, respectively. Bold type indicates the preferred model in each case. As predicted, a cover story consistent with a strong sampling assumption lead to non-monotonic responding in the **Target** trials, but not the **Control** trial, while a cover story consistent with a weak sampling assumption induced monotonic responding across all conditions and trials. Bayes factors are shown to two significant figures.

three hypotheses: that responding was monotonic (positive change: $\mu > 0$), non-monotonic (negative change: $\mu < 0$) or that the additional premise had no influence (null effect: $\mu = 0$). Analyses were conducted using the BayesFactor package in R (Morey & Rouder, 2014), applying the method outlined by Morey and Wagenmakers (2014) to test one-sided hypotheses. The results of these analyses are summarized in Table 3, which reports the Bayes factor between the two best hypotheses in each case. As the table makes clear, there is strong evidence for monotonic reasoning on the control trials regardless of condition, but there is evidence for a shift from monotonic to non-monotonic reasoning in the target conditions.

For the two conditions employing a neutral cover story, our intuition was that a mixed sampling assumption should be induced. Consequently, we expected mean response change in the Relevant Fillers and Random Fillers conditions to be within the bounds of that for the Both Relevant and Both Random conditions. To investigate this intuition, we determined the mix of strong and weak sampling assumptions (captured by θ, as per Equation (5)) that best fit the mean response change observed for each condition. The fitting process involved finding a value for θ (in the range 0 to 1) that minimised the squared difference between predicted response change and mean observed response change summed across **Control** and **Target** trials.

As Figure 3(b) shows, the change in relative mixture across conditions follows the expected pattern. The correlation between fitted model and data is 0.94, indicating a good fit overall. Further analysis showed that order restricted models suggesting either an effect of cover story only or both cover story and filler items were both well supported by the data, with the latter having strongest support overall (Bayes factors

		Bayes Factor (: NO EFFECT)		
Model	Order restrictions	Target 1	Target 2	Control
NO EFFECT	$\mu_1 = \mu_2 = \mu_3 = \mu_4$	-	-	-
FILLERS ONLY	$\mu_1 = \mu_2 < \mu_3 = \mu_4$	740:1	12,000:1	$< 1:1$
STORY ONLY	$\mu_1 < \mu_2 = \mu_3 < \mu_4$	4,100:1	17,000:1	$< 1:1$
BOTH	$\mu_1 < \mu_2 < \mu_3 < \mu_4$	2,900:1	30,000:1	$< 1:1$
RANDOM EFFECT	$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$	520:1	4,600:1	$< 1:1$

Table 4: Bayes factors representing the relative likelihood of the observed changes in argument strength under each model compared with the NO EFFECT model. A higher Bayes factor indicates greater evidence in favour of a particular model. Each model is described in terms of the order restrictions amongst the values μ_1, μ_2, μ_3 and μ_4, which represent the true means of the BOTH RELEVANT, RELEVANT FILLERS, RANDOM FILLERS, and BOTH RANDOM conditions, respectively. Bayes factors are shown to two significant figures.

are shown in Table 4). Bayes factors were calculated using a custom JAGS model, employing the product space method of model comparison (Lodewyckx et al., 2011; see Appendix B for details).

Overall, the effect of sampling assumption on premise monotonicity in our experiment was strong enough to cause a genuine reversal in whether people were prepared to endorse the conclusion in one case. For the second target trial 78% of participants in the BOTH RANDOM condition endorsed the conclusion that lions produce the hormone TH-L2, compared to 37% in the BOTH RELEVANT condition. With respect to the first target trial the effect was less pronounced due to low overall endorsement of the conclusion; 24% endorsement in the BOTH RANDOM condition compared to 11% in the BOTH RELEVANT condition.

Discussion

Arguments, when presented in everyday life, are intended to bring about a change in the audience. Whether to engage, to teach or persuade, premises are typically selected with a relevant goal in mind. This paper investigates why premise relevance should matter when people evaluate arguments. We demonstrate that people's reasoning in a category-based induction task is dependent on their assumptions about how the premises were sampled. If they think the premises were provided by a helpful confederate choosing positive examples from the categories in question, they show the premise non-monotonicity effect found previously (Medin et al., 2003). However, if they believe that the premises were generated randomly, this effect reverses. These results can be explained by a Bayesian theory of category-based induction that naturally incorporates different assumptions about premise sampling.

Our results support two qualitatively different conclusions. First, our work

shows that the perceived strength of an inductive argument is influenced not just by the direct generalizability of premises to conclusion, but also by expectations of premise relevance. By inducing a weak sampling assumption we showed that sensitivity to premise relationships can be violated. Second, this influence is pronounced enough to lead to a reversal of an effect (premise non-monotonicity) that normally obtains for certain kinds of argument structures. Reasoners who hold different sampling assumptions may endorse opposite conclusions as a result.

A previous attempt by Fernbach (2006) to demonstrate premise non-monotonicity by inducing a weak sampling assumption was not entirely successful. Although Fernbach (2006) found a difference in argument strength depending on sampling assumptions, participants in that study did not show a qualitative shift from monotonic to non-monotonic reasoning. Instead, the additional premises raised argument strength in all cases. It is possible that the relevance of the additional premises was not clear enough in that manipulation, which did not vary filler items. In our experiment we used filler items to substantiate the cover story in the Both Relevant and Both Random conditions. For example, our Both Random condition contained negative examples as filler items, without which a weak sampling assumption is difficult to sustain. Previous work involving category learning has also found that people rely on data, not just cover stories, to determine which sampling assumptions are appropriate. For instance, Navarro et al. (2012) found that the data people were shown affected their generalizations, but that sampling assumptions implicit in the cover story did not. A replication of that study which made the sampling assumptions in the cover story more explicit did find a reliable effect of cover story (Vong, Hendrickson, Perfors, & Navarro, 2013). Our results showed a reliable effect of both cover story and filler items, with participants in the Relevant Fillers and Random Fillers conditions exhibiting a similar, albeit attenuated, pattern of responding to those in the Both Relevant and Both Random conditions, respectively. This lends further support to the intuitive notion that in many cases people's sampling assumptions reflect some weighted mixture of strong and weak sampling. And while the questions remains open as to whether and how sampling assumptions are updated as new data arrives, it is clear that people do pay attention to the nature of the data when determining how that data was generated.

Prominent models of inductive argument strength, such as the similarity-coverage model of Osherson et al. (1990), and the featural similarity model of Sloman (1993) suggest that argument strength is based on the similarity between premise and conclusion, as first observed by Rips (1975). However, these models offer no explicit mechanism to capture sampling assumptions. Each model "hard-wires" a particular assumption instead. In contrast, as we have shown, a Bayesian model along the lines we have illustrated can accommodate the roles of both premise-conclusion similarity and sampling assumptions.

How might the relevance framework for inductive reasoning (Wilson & Sperber, 2004) accommodate our finding that premise sampling assumptions affect argument

strength? Relevance theory claims that an input is worth picking out from the mass of competing stimuli when it is *more* relevant, and that an input is more relevant if it produces a larger cognitive effect or requires less effort to process. The addition of a premise that highlights a shared property should raise the relevance of that property when determining the appropriate basis for induction, by decreasing the effort required to call the property to mind. But that should be so in each of our experimental conditions, because identical premises were used in the trials of interest. So that leaves us to posit a difference in cognitive effect to explain a difference in relevance between conditions.

This is where the Bayesian theory of category-based induction comes in. The theory describes how beliefs are revised in response to evidence in terms of the redistribution of probability mass. Such redistribution, we argue, is an excellent candidate measure for cognitive effect. Under this view, the mathematics of Bayes' rule predicts that a strong sampling assumption will always lead to a greater cognitive effect than would a weak sampling assumption because it leads to belief revision due to differences in the likelihood of observing certain data, and not simply due to falsification alone.[5] Relevance theory holds that comparing stimuli on the basis of relevance is a crucial part of human reasoning. The Bayesian theory of category-based induction provides a computational basis for making such comparisons in a way that takes two critical factors – premise sampling assumptions and category similarity – into account. As such, the theory represents an important component that can be integrated into the relevance framework. Likewise, relevance theory complements Bayesian theory insofar as it can make qualitative predictions regarding processing effort. Any algorithmic account of category-based induction should take these predictions into account, as well as relevant empirical findings (e.g. Coley & Vasilyeva, 2010; Feeney, Coley, & Crisp, 2010; Feeney & Heit, 2011).

In general, we found that people in our experiment quite naturally assumed that premises were selected sensibly or drawn from the category – the difficulty came in trying to persuade them that they were truly random, as in the Both Random condition. This observation, in combination with the fact that the premise non-monotonicity found in the Both Relevant condition corresponds to the standard effect (Medin et al., 2003), suggests that people have an automatic bias to believe that premises are selected sensibly: if not by a helpful teacher, at least in a way consistent with strong sampling (i.e., selected from the category). A biased presumption of relevance is an outcome in keeping with a central claim of relevance theory that people act to maximise relevance when selecting inputs to process (Wilson & Sperber, 2004). This is sensible in the context of category-based induction given that this is how arguments are constructed and used in the real world, but it does mean that we cannot, as researchers, assume that people reason as if we are generating examples

[5] An important implication of this assumption is that equating cognitive effect directly with change in argument strength is potentially flawed, since the two forms of belief revision can have opposing effects.

randomly (even when we are).

It should be noted that our model incorporates strong sampling, which in the context of category-based induction implies that a category exhibiting the property in question is as likely as any other to be chosen. Seeking to persuade or dissuade another is typically a matter of picking a relevant example of a concept, not a random one. Yet, when a property defines a small or coherent category such as "species of bear" or "black and white striped animals" then there is likely to be little variation in relevance across the category members, and a strong sampling assumption may be appropriate. A pedagogical assumption, in contrast, which gives greater weight to examples that better characterise a property, may be more appropriate for larger, less coherent categories, where there is greater variation in relevance across category members.[6] Shafto et al. (2014) found evidence to suggest that pedagogical sampling compared to strong sampling lead to tighter generalizations on the part of the learner, albeit with simple perceptual stimuli. It is plausible that our Both Relevant cover story acted to tighten generalizations over and above the predictions of strong sampling. Such a tightening may have acted to increase levels of premise non-monotonicity in the Both Relevant condition. Further work is needed to determine whether premise non-monotonicity can be observed with a cover story suggestive of a strong sampling assumption alone, in line with our model simulations. Regardless, the likelihood function in the Bayesian model may be adapted to capture either strong or pedagogical sampling (Shafto et al., 2014).

There is substantial evidence to suggest that when attempting to learn, generalize and draw conclusions from data, people are sensitive to the process by which data is generated. This sensitivity to sampling has been previously shown in simple generalization problems (Tenenbaum & Griffiths, 2001; Navarro et al., 2012), in early word learning (Xu & Tenenbaum, 2007), and even in infants (Gweon, Tenenbaum, & Schulz, 2010). Other work has demonstrated that people are sensitive to more complicated sampling schemes (Shafto et al., 2014). Our work extends this sensitivity to category-based induction tasks, adding an important clarification to relevance theoretic accounts of a phenomena attributed to relationships amongst premise items. In a world of exclusively weak sampling assumptions, where evidence supports falsification only, the inferential leap receives no boost from premise relevance: the relevant becomes irrelevant.

Appendix A

In order to generate model predictions (using Equations (1) to (4) described in the main paper) it is necessary to specify an hypothesis space $\mathcal{H}$ and a prior distribution, $P(\mathcal{H})$. To do so, we restrict the category labels under consideration to the fourteen experimental stimuli used in both experimental conditions. This is not to

[6] Pedagogical sampling (Shafto et al., 2014) may be viewed as a partial instantiation of the *communicative principle of relevance* (Wilson & Sperber, 2004), insofar as it can make predictions about belief revision in an explicitly communicative context.

say that the experimental participants were aware in advance of the nature and extent of the stimuli used, nor restricted their considerations in this manner. We made this restriction to render analysis tractable, with the view that the predictions remain valid in a qualitative sense, despite this truncation. The fact that our experimental results match our predictions in qualitative terms lends support to this view. Given the fourteen category labels, our hypothesis space $\mathcal{H}$ consists of 2^{14} hypotheses, each corresponding to the proposition that a unique cluster of categories share a given property.

Having established our hypothesis space $\mathcal{H}$, we need to separately derive a plausible prior distribution, $P(h)$, defined over all $h \in \mathcal{H}$. We seek a prior that is independent of any particular property or this specific task, to avoid fitting our predictions too tightly to the properties used in our experimental trials. That is, $P(h)$ represents the probability that a blank (unseen) property is shared by those items that belong to a particular category h. In keeping with prominent models of category-based induction (Osherson et al., 1990; Sloman, 1993), we assume that generalizing a property from one item to another involves an assessment of their similarity. Intuitively, since hypotheses in our model correspond to clusters of items, we seek to establish a weighting for each cluster that reflects its coherence. Prior probabilities will be derived from these clusters, with higher prior probabilities assigned to more coherent clusters.

To establish clusters and associated weights we apply the *additive clustering* (ADCLUS) model (Shepard & Arabie, 1979; Lee, 2002; Navarro & Griffiths, 2008) to similarly data gathered from a separate experiment, described in more detail below. On the basis of observed similarity data, ADCLUS identifies structure in the domain free from the undesirable restriction that such structure take a strictly hierarchical form. The model defines the similarity of any two objects as the sum of the weights across all clusters containing both objects. It attempts to find a set of clusters and weights maximising the fit between empirical similarity data and the theoretically reconstructed measures. Finding an optimal fit is an under-constrained and computationally expensive exercise, hence the model implementation seeks to find a good and parsimonious fit. Starting with an initial configuration of clusters and weights, a gradient descent algorithm is employed to find a suitable local optimum. On each iteration of the gradient descent process, clusters with non-appreciable weights may be discarded.

In order to provide empirical interstimulus similarities as input to the ADCLUS model, a separate experiment was conducted to gather similarity ratings via a triad task for the fourteen animal stimuli common to all conditions of our experiment. 63 adults were recruited via Amazon Mechanical Turk, and were each paid \$0.60 (USD) for the 5–10 minutes participation. 5 were excluded due to browser incompatibility, and the remaining 58 were aged 19 to 75 years (median age 31, 41% female). 50 participants were in the United States, with 8 located elsewhere. For each triad presented, people were asked to pick which animal was *least* similar to the others. Each

person rated 60 randomly selected triads. Since there were a total of 364 possible triads, this meant that each triad was rated by 9–10 participants on average. The pairwise interstimulus similarity for two stimuli a and b was calculated as the proportion of all triad ratings for a, b, and some other stimulus c, where c was rated as being the least similar.

The final stage in our model implementation involves the assignment of prior probabilities based on the clusters and weights identified by ADCLUS. Let $\mathcal{H}_C$ denote those hypotheses (clusters) identified by the ADCLUS process, and w_h denote the weight associated with hypothesis $h \in \mathcal{H}_C$. We form an initial estimate of the prior distribution directly from these outputs:

$$P_w(h) \propto \begin{cases} w_h & \text{if } h \in \mathcal{H}_C \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

This initial estimate is not quite right, however. The ADCLUS model does not deal meaningfully with clusters corresponding to a single category. Yet intuitively, in the context of our experiment, properties that pertain to a single category (TIGER, for example) are quite plausible. Therefore we need to combine the prior derived from the cluster weights with one that assigns non-zero probability to the singleton hypotheses (the set of which we denote $\mathcal{H}_S$). For the latter, we use a size-based prior:

$$P_s(h) \propto \begin{cases} \frac{1}{|h|} & \text{if } h \in \mathcal{H}_C \cup \mathcal{H}_S \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Lastly, we combine these two prior distributions to form the prior used to generate our model predictions in such a way that the probabilities for singleton hypotheses calculated in Equation (7) are preserved:

$$P(h) = \begin{cases} P_s(h) & \text{if } h \in \mathcal{H}_S \\ P_w(h) \sum_{h' \in \mathcal{H}_C} P_s(h') & \text{if } h \in \mathcal{H}_C \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

As the reader will note, our method for defining the hypothesis space and for deriving prior probabilities affords a certain latitude. Using the ADCLUS model, the precise clusters and associated weights identified depend on the values chosen to seed the optimization process. Whilst we retain the seeding heuristic of Shepard and Arabie (1979), we also experimented with other heuristics. We found that although such alternatives lead to different numerical predictions, the important qualitative effect was robust: greater levels of premise non-monotonicty were predicted under a strong sampling assumption than under a weak sampling assumption for the **Target** (but not the **Control**) arguments. Similarly, alternative methods may be employed for assigning probabilities to singleton hypothesis, but once again, the qualitative predictions appear robust in the face of such changes.

Appendix B

As discussed in the main text, differences in mean change in argument strength across conditions indicated that our experimental manipulation had some effect. To investigate the factors driving the effect we compared a number of plausible models to determine which might best account for our experimental results. The models considered were based on the change in argument strength predicted by our Bayesian model of category-based induction, derived from empirical similarity ratings. Under a strong sampling assumption, our model predicts non-monotonic responding for both **Target** trials; under a weak sampling assumption, monotonic responding is predicted.

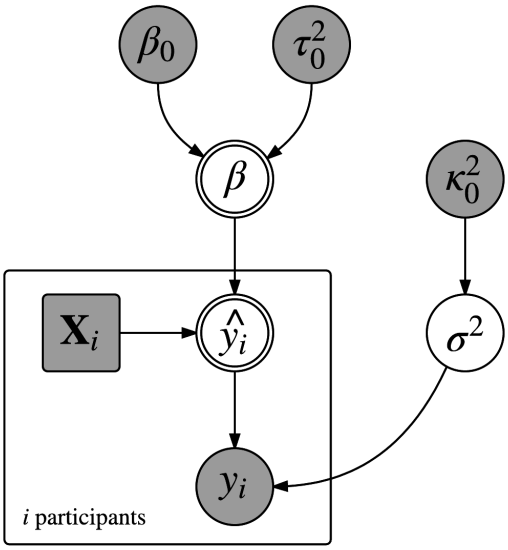
Furthermore, the fitted values of θ derived from the mixed sampling model suggest an ordering in terms of mean response change across conditions. Thus, consistent with the suggested orderings, three plausible models concerning the nature of the effect were compared, namely: that the effect was driven by the filler items only (FILLERS ONLY), that it was driven by the cover story only (STORY ONLY), or that it was driven by both of these factors (BOTH). The order restrictions for each model are shown in Table 5. A fourth unrestricted model was also considered, namely that results were driven by a random effect (RANDOM EFFECT).

For each of the four models, we calculated the Bayes factor representing the relative likelihood of the observed changes in argument strength under the model against the "no effect" model (NO EFFECT). To do so, we employed a Markov chain Monte Carlo (MCMC) procedure known as the *product space method* (Lodewyckx et al., 2011). The technique supports the comparison of two models (M_o and M_1, for example) by building a hierarchical "supermodel" combining the models via a random variable (M, say) that acts as a model index. The Bayes factor for the relative likelihood of M_1 against M_0 becomes the posterior odds ratio ($M_1 : M_0$) for the two models, divided by the prior odds ratio. Theoretically, the prior model probabilities may be chosen with freedom, although technical considerations require careful selection if reliable MCMC estimates are to be obtained. Finally, the prior probabilities for each model may be estimated as follows:

$$\hat{P}(M_k \,|\, \text{Data}) = \frac{\text{Number of posterior samples where } M = k}{\text{Total number of posterior samples}}, \tag{9}$$

from which the Bayes factor easily follows.

Figure 4 shows the graphical model capturing the common elements for each of the models tested. The vector quantity $\boldsymbol{X_i} = (X_{1i}, X_{2i}, X_{3i})$ represents a dummy coding of condition for each participant. The vector quantity $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ captures the relationship between the means μ_1, μ_2, μ_3, and μ_4 of the BOTH RELEVANT, RELEVANT FILLERS, RANDOM FILLERS, and BOTH RANDOM conditions, respectively; that is, $\beta_1 = \mu_1$, $\beta_2 = \mu_2 - \mu_1$, $\beta_3 = \mu_3 - \mu_1$, and $\beta_4 = \mu_4 - \mu_1$. The δ_i parameters represent the difference between adjacent condition means, and are each sampled from a normal distribution with mean 0 and variance τ_0^2. The range restrictions on the values sampled differ across the five models, as shown in Table 5. The



$$
\begin{aligned}
y_i &\sim \text{Normal}\big(\hat{y}_i, \sigma^2\big) \\
\hat{y}_i &= \beta_1 + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 X_{3i} \\
\sigma^2 &\sim \text{Scaled Inverse } \chi^2\big(1, \kappa_0^2\big) \\
\beta_1 &\sim \text{Normal}\big(\beta_0, \tau_0^2\big) \\
\beta_2 &= \delta_1 \\
\beta_3 &= \delta_1 + \delta_2 \\
\beta_4 &= \delta_1 + \delta_2 + \delta_3
\end{aligned}
$$

Figure 4. A graphical model supporting comparison of condition means. For each of the five models considered, β_1 represents the condition mean of the reference condition Both Relevant. β_2, β_3, and β_4, represent the difference between the mean of the reference condition and the mean of the Relevant Fillers, Random Fillers, and Both Random conditions, respectively. The models differ only in the definition of δ_1, δ_2, and δ_3.

mean of the reference condition has a normal prior distribution with mean β_0, and variance τ_0^2. The prior for the error variance (σ^2) is a scaled inverse χ^2 distribution, with 1 degree of freedom and scaling parameter κ_0^2. To ensure that these prior distributions do not favour any one particular model, and that the posterior is effectively independent of the prior, the values for β_0, τ_0^2, and κ_0^2 were derived from the data using the procedure outlined in Klugkist, Laudy, and Hoijtink (2005, p. 482).

References

Coley, J. D., & Vasilyeva, N. Y. (2010). Chapter 5 - generating inductive inferences: Premise relations and property effects. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 53, pp. 183–226). Academic Press.

Feeney, A., Coley, J. D., & Crisp, A. K. (2010). The relevance framework for category-based induction: Evidence from garden-path arguments. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *36*(4), 906–919.

Feeney, A., & Heit, E. (2011). Properties of the diversity effect in category-based inductive reasoning. *Thinking & Reasoning*, *17*(2), 156–181.

Fernbach, P. M. (2006). Sampling assumptions and the size principle in property induction. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1287–1293). New York: Psychology Press.

		Parameter range		
Model	Order restrictions	δ_1	δ_2	δ_3
NO EFFECT	$\mu_1 = \mu_2 = \mu_3 = \mu_4$	0	0	0
FILLERS ONLY	$\mu_1 = \mu_2 < \mu_3 = \mu_4$	0	$(0, \infty)$	0
STORY ONLY	$\mu_1 < \mu_2 = \mu_3 < \mu_4$	$(0, \infty)$	0	$(0, \infty)$
BOTH	$\mu_1 < \mu_2 < \mu_3 < \mu_4$	$(0, \infty)$	$(0, \infty)$	$(0, \infty)$
RANDOM EFFECT	$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$

Table 5: The range restriction imposed on the Normal$\left(0, \tau_0^2\right)$ distribution from which the δ_i parameters are sampled for each model. A value of 0 indicates that the respective parameter is always 0.

Grice, H. P. (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066–9071.

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.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20-58.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: a Bayesian approach. *Psychological Methods*, *10*(4), 477–493.

Lee, M. D. (2002). Generating additive clustering models with minimal stochastic complexity. *Journal of Classification*, *19*(1), 69–85.

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*(5), 331 - 347. doi: http://dx.doi.org/10.1016/j.jmp.2011.06.001

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin and Review*, *10*(3), 517–532.

Morey, R. D., & Rouder, J. N. (2014). BayesFactor: Computation of Bayes factors for common designs [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.9)

Morey, R. D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics and Probability Letters*, *92*, 121-124.

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223.

Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural computation*, *20*(11), 2597–2628.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120.

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

Rips, L. (1975). Inductive judgments about natural categories. *Verbal Learning and Verbal Behaviour*, *14*, 665–681.

Sanjana, N. E., & Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (pp. 59–66). MIT Press.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87–123.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 213–280.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Vong, W. K., Hendrickson, A. T., Perfors, A., & Navarro, D. J. (2013). The role of sampling assumptions in generalization with multiple categories. In *Proceedings of the 35th annual meeting of the cognitive science society* (pp. 3699–3704).

Voorspoels, W., Van Meel, C., & Storms, G. (2013). Negative observations, induction and the generation of hypotheses. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1552–1557).

Wilson, D., & Sperber, D. (2004). Relevance theory. In L. Horn & G. Ward (Eds.), *Handbook of pragmatics* (pp. 607–632). Oxford: Blackwell.

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297.