How many kinds of reasoning? Inference, probability, and natural language semantics

Daniel Lassiter, Noah D. Goodman
Department of Psychology, Stanford University
{danlassiter, ngoodman} @ stanford.edu

Abstract

Previous research (Heit & Rotello, 2010; Rips, 2001; Rotello & Heit, 2009) has suggested that differences between inductive and deductive reasoning cannot be explained by probabilistic theories, and instead support two-process accounts of reasoning. We provide a probabilistic model that predicts the observed non-linearities and makes quantitative predictions about responses as a function of argument strength. Predictions were tested using a novel experimental paradigm that elicits the previously-reported response patterns with a minimal manipulation, changing only one word between conditions. We also found a good fit with quantitative model predictions, indicating that a probabilistic theory of reasoning can account in a clear and parsimonious way for qualitative and quantitative data previously argued to falsify them. We also relate our model to recent work in linguistics, arguing that careful attention to the semantics of language used to pose reasoning problems will sharpen the questions asked in the psychology of reasoning.

Keywords: Reasoning, induction, deduction, probabilistic model, formal semantics.

Suppose that you have learned a new biological fact about mammals: whales and dogs both use enzyme B-32 to digest their food. Is it now *necessary* that horses do the same? Is it *plausible*, *possible*, or *more likely than not*? Expressions of this type—known as *epistemic modals* in linguistics—have played a crucial role in recent work that argues for a sharp qualitative distinction between inductive and deductive modes of reasoning. In the paradigm introduced by Rips (2001) and extended by Heit and Rotello (2010); Rotello and Heit (2009), participants are divided into two conditions and are either asked to judge whether a conclusion is "necessary" assuming that some premises are true, or whether it is "plausible". The former is identified with the deductive mode of reasoning, and the latter with the inductive mode.

These authors asked participants in both conditions to evaluate a variety of logically valid and logically invalid arguments. An example invalid argument might be "Cows have sesamoid bones; Mice have sesamoid bones; therefore, Horses have sesamoid bones". An example valid argument might be "Mammals have sesamoid bones; therefore, horses have sesamoid bones." They found that there was a non-linear relationship between the endorsement rates of arguments depending on condition: participants in both conditions generally endorsed logically valid arguments, but participants in the deductive condition were much less likely to endorse invalid arguments than those in the inductive condition. These results are interpreted as a challenge to theories of reasoning which rely on a single dimension of argument strength and interpret deductive validity as simply the upper extreme of this dimension(Harman, 1999; Johnson-Laird, 1994; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990). In particular, Rips and Heit & Rotello argue that non-linearities cannot be accounted for by probabilistic theories of reasoning, which identify the strength of an argument with the conditional probability of the conclusion given the premises (Heit, 1998; Kemp & Tenenbaum, 2009; Oaksford & Chater, 2007; Tenenbaum, Griffiths, & Kemp, 2006). On the other hand, they claim that the results are consistent with two-process theories of reasoning (Evans & Over, 1996).

We argue that the manipulation involving "necessary" and "plausible" hinges not on a qualitative distinction between two reasoning processes, but rather on facts about the semantics of these words which can be modeled using a single underlying scale of argument strength—conditional probability. We propose a semantically motivated model of reasoning with epistemic concepts which predicts non-linear response patterns depending on the choice of modal similar to those observed in previous work, and makes detailed quantitative predictions about response patterns in invalid arguments.

We test the claim that the modal word is the crucial factor using a new paradigm that isolates its effects. Our arguments had the same form as the examples above, except that we placed the modal word of interest in the conclusion:

Premise 1: Cows have sesamoid bones.
Premise 2: Mice have sesamoid bones.
Conclusion: It is {plausible/necessary/possible/likely/probable/certain} that horses have sesamoid bones.

We will refer to configurations such as "It is plausible/possible/etc. that *C*" as a **modal frame**. If varying the modal frame gives rise to a non-linear pattern of responses similar to the one found in previous work, this would indicate that an explanation of these results should be framed in terms of the meaning of these modal words.

Together, the model and experimental evidence indicate that the negative conclusions of previous work regarding one-dimensional theories of argument strength are not warranted: it is possible to explain non-linear response patterns with a probabilistic account of argument strength.

Previous Work

Rips (2001) conducted a reasoning experiment designed to investigate the traditional distinction between deductive and inductive reasoning. Participants in two groups were asked to judge arguments either according to whether the conclusion was *necessary* (assuming that the premises were true) or whether it was *plausible*. Most participants in both conditions accepted logically valid arguments and rejected invalid arguments whose conclusion was not causally consistent with the premises, such as "Car X strikes a wall, so Car X speeds up". However, participants differed by condition in whether they rejected non-valid arguments which were causally consistent with the premises: those in the inductive condition generally accepted arguments such as "Car X strikes a wall, so Car X slows down", while those in the deductive condition did not. Rips argued that this result falsifies theories of reasoning in which argument strength is a one-dimensional quantity such as conditional probability: "[i]f participants base all forms of argument evaluation on the position of the argument on a single psychological dimension, then induction and deduction judgments should increase or decrease together" (p.133).[1]

Heit and Rotello (2010); Rotello and Heit (2009) extended Rips' paradigm in a number of ways. Their core finding was that d', a standard measure of sensitivity in Signal Detection Theory (SDT), was significantly higher in the deductive condition across a variety of arguments types and manipulations. d' is defined as $z(H) - z(F)$, the difference between the z-scored hit rate H and false alarm rate F (Macmillan & Creelman, 2005). This difference means that participants in the inductive condition were more sensitive to argument validity than participants in the deductive condition (see Table 1).

Table 1: Acceptance rates and d' in Experiment (1a) of Rotello and Heit (2009) (three-premise arguments only).

	Deduction	Induction
Acceptance, valid	.94	.95
Acceptance, invalid	.06	.17
Sensitivity (d')	3.31	2.56

Differential sensitivity indicates that the difference between conditions is not simply a shift in response criterion in the presence of two equal-variance Gaussians representing signal and noise. Thus we cannot fit a one-dimensional SDT model to such results. In accord with Rips, Rotello and Heit (2009) argue that the non-linear relationship between validity and condition is a challenge to probabilistic theories of reasoning. They argue that the results are better captured by a two-dimensional SDT model with possibly orthogonal dimensions of inductive strength and deductive validity, in which the response criterion can vary in two dimensions.[2]

A Probabilistic Model

In contrast to Rips and Rotello & Heit, we do not see these results as strong support for a two-process theory. Instead, we will argue that these results are also compatible with a probabilistic account of inductive reasoning once the semantics of the modal words used in the experiment is taken into account. In this section we propose a model of the relationship between the probability on the one hand and epistemic concepts such as certainty, necessity, and plausibility on the other. The latter are treated as non-linear functions of the former determined by a single parameter per item. This model, inspired by recent work in formal semantics, predicts non-linear response patterns and variation in d' depending on the modal used, and makes a number of fine-grained predictions that we will later evaluate against the results of an experiment.

Our probabilistic model of reasoning with epistemic modals is intended to capture the following intuitions. A maximally strong conclusion C remains maximally strong whether you ask if it is *possible, plausible, likely* or *necessary*; a maximally weak conclusion remains maximally weak under the same conditions; but there is much more flexibility around the middle of the probability scale depending on which question is asked. If C has a probability of .4, it presumably will count as *possible* and perhaps as *plausible*, but it would not seem to be *likely* and surely not *necessary* or *certain*. Thus the effect of an epistemic modal on a conditional probability should be a transformation that preserves the minimum value 0 and the maximum value 1.

Perhaps the simplest way to capture the behavior just described is to suppose that each modal $M \in \{$*possible, plausible, likely, probable, certain, necessary*$\}$ is associated with a parameter $\alpha_M \in \mathbb{R}^+$ which, in combination with the conditional probability $pr(C|P)$ of conclusion C given premises P, determines the probability of *It is M that C* given P. We propose that α_M relates these two probabilities by a power-law:

$$pr(\textit{It is M that } C|P) = pr(C|P)^{\alpha_M} \tag{1}$$

We model arguments with no modal as also being governed by some some α_M as in (1) (rather than directly reflecting the conditional probability of the conclusion given the premises).

We assume that participants' responses to a modalized question Q are governed by (1) plus a noise parameter ε (interpreted as the proportion of trials in which participants choose a response at random).

$$pr(\text{"yes"}|P, Q = \textit{Is it M that C?}) = pr(C|P)^{\alpha_M}(1-\varepsilon) + \frac{\varepsilon}{2} \tag{2}$$

Depending on α_M we get a variety of possible curves, a few of which are sketched in figure 1b (setting $\varepsilon = .1$ for illustrative purposes). This model captures the behavior just described: in the limits α_M does not influence the response probability, but there is significant variation in response probabilities in the middle range depending on α_M. This feature leads to a prediction that that the choice of modal will have less influence on response rates for arguments with very high strength (e.g., valid arguments) than those with intermediate strength.

[1] Rips (2001) also found a crossover effect in which participants in the "necessary" condition were slightly more likely to endorse valid arguments that were inconsistent with causal knowledge than participants in the "plausible" condition, but the reverse was true for arguments consistent with causal knowledge. Heit & Rotello's work did not find any analogous effect, nor did we in our experiment reported below, and we do not consider it further. However, it is possible that the effect is real and attributable to one of the various detailed differences in materials and instructions between the experiments.

[2] Heit and Rotello (2010); Rotello and Heit (2009) also introduced a number of further manipulations involving e.g. constrained response time, number of premises, and readability which we will not discuss in detail for reasons of space. See the concluding section for a brief consideration of two of these manipulations, however.

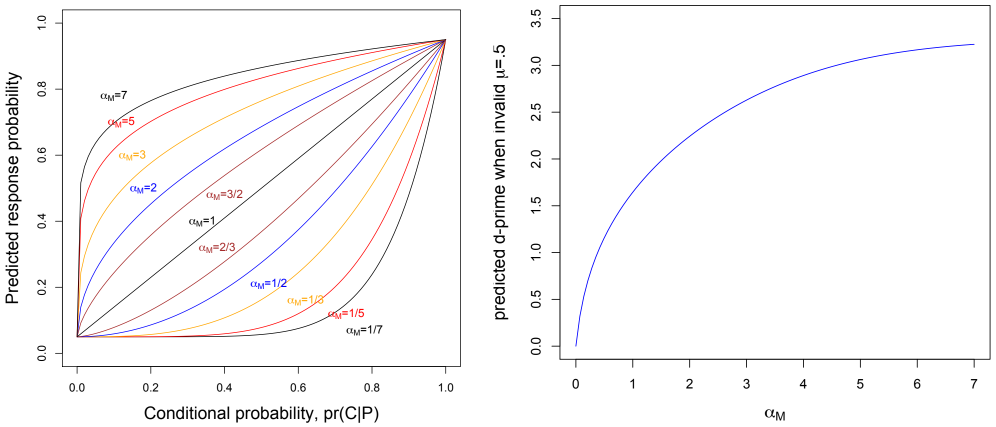

Figure 1: (**a**): Predicted response probability for various settings of α_M. (**b**): Example of relation between α_M and d'.

We also predict that d' will vary depending on α_M. Suppose for illustration that the mean conditional probability of logically invalid arguments in some sample is .5, and that the conditional probability of valid arguments is 1. The d' statistic estimated from this data should then be (on average)

$$\begin{aligned} d' &= z(1^{\alpha_M}(1-\varepsilon)+\varepsilon/2)-z(.5^{\alpha_M}(1-\varepsilon)+\varepsilon/2) \\ &= z(1-\varepsilon/2)-z(.5^{\alpha_M}(1-\varepsilon)+\varepsilon/2) \end{aligned} \quad (3)$$

If $\varepsilon = .1$, we expect the observed d' to be related to α_M as in figure 1b. This illustrates the fact that the value of the d' statistic is not predicted to be constant in our probabilistic model, but should depend on the choice of M. Thus, a model with one dimension of argument strength (conditional probability) is able to predict non-linearities of the type previously claimed to be problematic for probabilistic accounts.

The model also makes strong quantitative predictions about the relationship between different modal frames. That is, we predict a systematic (though non-linear) relationship between the response rates to the same argument in the modal frames *It is* M_1 *that* C and *It is* M_2 *that* C. If M_1 is associated with parameter α_1 and M_2 with α_2, for any argument with premises P and conclusion C there is some positive r such that

$$\begin{aligned} pr(\textit{Is it } M_1 \textit{ that } C|P) &= pr(C|P)^{\alpha_{M_1}} \\ &= pr(C|P)^{(r\times\alpha_{M_2})} \\ &= pr(\textit{Is it } M_2 \textit{ that } C|P)^r \end{aligned} \quad (4)$$

The prediction that every pair of modals should be related by a power-law allows us to evaluate model fit using a variety of arguments which, although not logically valid, vary widely in intuitive strength. It also shows that our model predicts that the strength of any two arguments should be related monotonically (though non-linearly) across modal frames.

Experiment

Our experiment tested the hypothesis that non-linear response patterns can be attributed to the semantics of the modal expressions used. The main innovation was to manipulate the choice of modal M within the stimulus sentence:

Non-valid:	*Valid*:
Cows have enzyme X.	Horses have enzyme X.
Seals have enzyme X.	Cows have enzyme X.
So, it is *M* that horses	So, it is *M* that horses
have enzyme X.	have enzyme X.

This minimal manipulation allowed us to isolate the effect of the modal frame on acceptance rates both for valid vs. invalid arguments and for invalid arguments of varying strengths. We also used a larger set of epistemic modal expressions than were used in previous work, including *possible, plausible, likely, probable, necessary*, and *certain*.

Methods

Participants 507 participants were recruited using Amazon's Mechanical Turk platform, with a restriction to participants located in the United States. They were compensated for their participation.

Materials Participants saw 21 valid and invalid arguments with two premises and a conclusion. 18 of these included one of the 6 modal words listed above (3 for each modal), and 3 did not include a modal word. The properties used in the arguments were chosen from a list of unfamiliar biological and pseudo-biological properties. In every case, the animal in the conclusion was "horses" (Osherson et al., 1990). The arguments seen were randomly selected from a total of 63 arguments tested. 36 were non-valid arguments, in which the premise animals were chosen from the set {*cows, chimps, gorillas, mice, squirrels, elephants, seals, rhinos, dolphins*}. 27 were logically valid arguments with one of two forms, following (Rotello & Heit, 2009): one premise involved a mammal and the other was an *identity* premise (stating that horses have the property) or an *inclusion* premise (stating that mammals or animals have the property).

Procedure Participants were instructed to answer each question according to whether they agreed with the conclusion, assuming that the premises were true. For each question participants were asked to select "agree" or "disagree" and to give a confidence rating on a five-point scale.

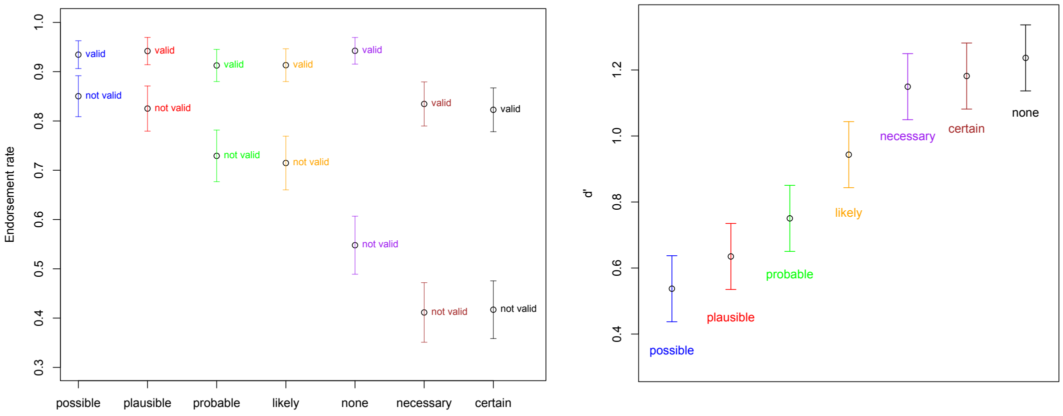

Figure 2: **(a)**: Valid and invalid endorsement rates by modal frame with 95% CIs; **(b)**: d' by modal frame with 95% CIs.

Results

We analyzed data from the 485 participants who reported that their native language was English, for 10,185 total judgments. No systematic differences emerged between identity and inclusion arguments, and we group them together as valid.

Our results replicated the crucial finding of (Heit & Rotello, 2010; Rotello & Heit, 2009) showing that sensitivity to argument validity is greater when participants are asked to judge whether a conclusion is "necessary" than when they are asked whether it is "plausible" (Table 2). The difference in d' values is significant at the $p < .05$ level.

Table 2: Acceptance rates and d' in our experiment (*plausible* and *necessary* only), with 95% confidence intervals.

	Necessary	*Plausible*
Acceptance, valid	.82	.94
Acceptance, invalid	.41	.82
Sensitivity (d')	1.15 ± .19	0.63 ± .27

Comparing Table 1 with Table 2, there are two clear differences between our results and those of Rotello and Heit (2009): our participants rejected valid arguments with *necessary* more often than with *plausible*, and they accepted invalid arguments at much higher rates. These factors contributed to lower d' in both conditions. The first difference suggests that participants in our experiment judged some logically valid arguments as strong but not maximally strong. The second is plausibly due to the use of different materials in our experiment: Rotello & Heit used the same predicate "have property X" in all arguments, using the variable X in the stimuli and instructing participants to treat property X as a novel biological property. The use of more natural biological predicates in our experiment may have encouraged participants to have greater confidence in their guesses, particularly if they had general background knowledge about e.g. enzymes and bones.

The fact that d' differed significantly for "necessary" and "plausible" suggests that our within-participants manipulation successfully captured the core features of between-participants manipulations in previous work. That is, the difference previously reported can be elicited by a difference of a single modal word, and so appears to be triggered by semantic properties of the modal words.

Assessing Model Fit As pointed out in introducing the model, our account predicts three general trends. First, it predicts that acceptance rates should vary less across valid arguments than across invalid arguments (Figure 1a). This is indeed what we find (Figure 2a). Second, it predicts the possibility of a continuous gradient in d' values (Figure 1b). Our results confirm this prediction as well (Figure 2b). Third, it predicts that the acceptance rate of argument A in any two modal frames M_1 and M_2 should be related by a power-law. This is the quantitative prediction that we now set out to test.

We fit the model to our results using the acceptance rates of each argument in the no-modal condition as a baseline. Note that we do not believe that acceptance rates in this condition are an estimate of the true probability of the conclusion given the premises. However, our model predicts that the choice of baseline condition should not affect predictions (cf. equation 4). In accord with this prediction we found no systematic effect on R^2 by varying the choice of baseline. The primary effect of the choice of baseline condition, then, is that estimates of α_M given for the other conditions are up to multiplication by α_0, the parameter which determines the probability of arguments in the baseline condition.

Figure 3 plots the endorsement rate of each argument in the unmodalized condition against the endorsement rate for the same argument in various modal frames. We calculated the best-fit α_M for each condition. For each graph in Figure 3 the curve determined by equation (2) is superimposed, with the overall best-fit noise parameter $\varepsilon = .12$. As the R^2 values in Figure 3 and Table 2 show, this model captures much of the variance in the data, but not all of it.

In order to discern whether the remaining variance was due to systematic factors that our model does not capture, we performed a split-half reliability test, randomly dividing the data into equal-sized halves 10,000 times and testing the correlation between the two halves on the same measure that

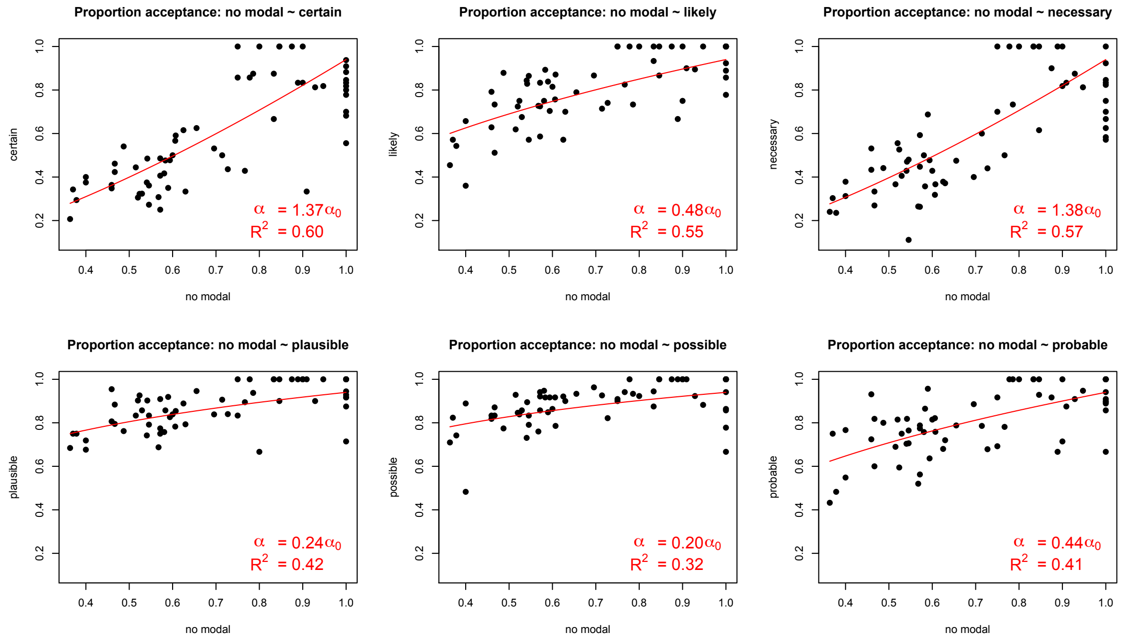

Figure 3: Endorsement rates by argument and modal frame. Curves represent model predictions using (2) and the best-fit α.

was used to fit our model: acceptance rate by argument and modal. The model correlations were overall very close to the split-half correlations (Table 3). This suggests that the model captures most of the true structure in the data, though a good deal of noise remains that we cannot expect to explain.

Recall that the model predicts a consistent order in the endorsement rates of arguments across modal frames (see equation 4). In order to test this prediction we considered two tests, a permutation test and a model-based simulation. The average value of Spearman's rank-order correlation for all pairwise comparisons between modal frames in our experiment was .53. A 10,000-sample permutation test revealed that this correlation was highly significant, $p < .0001$.

For the model-based test, we simulated 1,000 data sets with the same dimensions as our data using the best-fit model predictions as binomial proportions for each argument and modal frame. Equation 4 entails that all rank-order correlation in the model predictions are 1, and so this test gives us an opportunity to observe what correlations we should expect to see in a data set of this size if the prediction of a perfect underlying rank-order correlation is correct. The result was consistent with the observed rank-order correlation of .53, with 95% of the model-generated correlations falling between .49 and .64 (mean = .57). This result suggests that the model's monotonicity predictions fit the data well, providing further support for our claim that argument strength is based on a single scale which is manipulated by modal expressions.

Table 3: Model correlations vs. split-half reliability results.

Modal	model R^2	mean split-half R^2
certain	.60	.66
likely	.55	.57
necessary	.57	.70
plausible	.42	.37
possible	.32	.34
probable	.41	.47

Relation to Formal Semantics

Epistemic modality has been the subject of much investigation in formal semantics (Kratzer, 1991; Egan & Weatherson, 2011). This paper has been concerned with a subtype of epistemic modals whose grammatical features indicate that they are *gradable adjectives*. This means that, like other gradable adjectives, they can be used in comparative and degree modification structures: for example, conclusion C might be *more plausible* than C' given some evidence, or C might be *very likely* or *almost necessary*. This corresponds to the fact that person x can be *taller* than person y, and a glass of water might be *very large* or *almost full*. Gradable expressions are generally treated in formal semantics as functions which map objects to points on a *scale* and compare them to *threshold values* (Kennedy, 2007). Recent empirical and formal work on gradable epistemic modals suggests that a scalar analysis is appropriate for them as well, and that probability is a good choice for the underlying scale (Lassiter, 2010; Yalcin, 2010).

Although our model is superficially different from linguistic models, there is in fact a straightforward translation between the two. Linguistic models assume that the location of the threshold is typically uncertain, a fact which is closely related to the problem of vagueness. If the uncertain location of the threshold for gradable expressions is modeled probabilistically as in (Frazee & Beaver, 2010; Lassiter, 2011; Schmidt, Goodman, Barner, & Tenenbaum, 2009), our model can be seen as describing the cumulative distribution of this threshold. Letting $p_M(\theta)$ be the probability density of the unknown threshold associated with modal M, the cumulative distribution is given by (5), which — plugging in (1) — gives us (6).

$$pr(\textit{It is M that } C|P) = \int_0^{pr(C|P)} p_M(\theta)\ \mathrm{d}\theta \qquad (5)$$

$$pr(C|P)^{\alpha_M} = \int_0^{pr(C|P)} p_M(\theta)\ \mathrm{d}\theta \qquad (6)$$

Using the fundamental theorem of calculus we can derive from equation (6) a simple formula linking $p_M(\theta)$ to α_M.

$$p_M(\theta) = \alpha_M \theta^{\alpha_M - 1}, \qquad 0 \leq \theta \leq 1. \qquad (7)$$

This relationship allows us to interpret our model as an implementation of scalar semantics for gradable epistemic modals closely related to recent work in formal semantics.

Conclusion

We have shown that non-linearities in responses to valid and invalid arguments can be explained by a simple probabilistic model, and thus are not evidence against a probabilistic account of reasoning. Our experimental manipulation involved the difference of a single word (an epistemic modal), and uncovered a gradation of non-linearities as a function of the specific modal used. Our model predicts a particular functional form for these differences, a power law in conditional probability, which we found in our data. This shows that it is possible to account for different patterns of results in reasoning experiments without assuming that everyday reasoning makes use of two (or more) qualitatively different types of reasoning. Rather, our model utilizes a single type of reasoning — probabilistic inference — together with a number of different but related linguistic mechanisms for talking about the results of inferential processes. This indicates that a one-dimensional theory of argument strength, coupled with an explicit formal semantics for epistemic modals, can account for a variety of patterns of reasoning in a parsimonious and explanatory way. This does not rule out a qualitative distinction between inductive and deductive reasoning, but it does call into question previous efforts to show that such a distinction is necessary to account for everyday human reasoning, suggesting instead that a single process may underlie both.

The probabilistic approach also suggests accounts of several related phenomena. For instance, the fact that argument length tends to affect plausibility judgments more than necessity judgments (Rotello & Heit, 2009) may be attributable to the fact that, in a probabilistic theory, we expect that adding premises of the type used in these experiments will usually increase the probability of the conclusion given the premises (Heit, 1998). The non-linearities predicted by equation 2 lead us to expect that the same change in probability will have different effects depending on the modal used.

Our probabilistic theory leaves open the psychological process by which people evaluate arguments. One possibility is that people are only able to *sample* possible worlds in accord with the distribution implied by the premises (Vul, Goodman, Griffiths, & Tenenbaum, 2009), and evaluate the truth of the conclusion in these sampled worlds. If people take several samples and respond "yes" when the conclusion is true in each sampled world, we recover a power law. If the average number of samples depends on the modal, we recover the probabilistic model described above. For instance, we would posit that people tend to take more samples to evaluate "necessary" conclusions than "plausible" conclusions. This process-level implementation predicts that under time pressure, when people can take fewer samples, "necessary" would begin to look more like "plausible". Indeed, this is exactly the finding of Heit and Rotello (2010).

We have illustrated an approach to reasoning based on an overall probabilistic view of inference, together with careful attention to natural language semantics. We believe that this approach will prove fruitful in studying a wide variety of phenomena related to human reasoning.

Acknowledgements

We thank Evan Heit for helpful discussion of an earlier draft. This work was supported by a John S. McDonnell Foundation Scholar Award and ONR grant N00014-09-1-0124.

References

Egan, A., & Weatherson, B. (2011). *Epistemic modality*. OUP.

Evans, J., & Over, D. (1996). *Rationality and reasoning*. Psychology Press.

Frazee, J., & Beaver, D. (2010). Vagueness is rational under uncertainty. *Proceedings of the 17th Amsterdam Colloquium*.

Harman, G. (1999). *Reasoning, meaning, and mind*. OUP.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. *Rational models of cognition*, 248–274.

Heit, E., & Rotello, C. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3).

Johnson-Laird, P. (1994). Mental models and probabilistic thinking. *Cognition*, *50*(1-3), 189–209.

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

Kennedy, C. (2007). Vagueness and grammar. *Ling. & Phil.*, *30*(1).

Kratzer, A. (1991). Modality. In A. von Stechow & D. Wunderlich (Eds.), *Semantics: Int'l hbk. of contemp. research.* de Gruyter.

Lassiter, D. (2010). Gradable epistemic modals, probability, and scale structure. In *Semantics and linguistic theory (SALT) 20.*

Lassiter, D. (2011). Vagueness as probabilistic linguistic knowledge. In Nouwen et al. (ed.)*, Vagueness in communication.* Springer.

Macmillan, N., & Creelman, C. (2005). *Detection theory: A user's guide*. Lawrence Erlbaum.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

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

Rips, L. (2001). Two kinds of reasoning. *Psych. Science*, *12*(2).

Rotello, C., & Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. *J. Exp. Psych.: Learning, Memory, and Cognition*, *35*(5).

Schmidt, L. A., Goodman, N. D., Barner, D., & Tenenbaum, J. B. (2009). How tall is *Tall*? Compositionality, statistics, and gradable adjectives. In *Proc. 31st annual CogSci.*

Tenenbaum, J., Griffiths, T., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Vul, E., Goodman, N., Griffiths, T., & Tenenbaum, J. (2009). One and done? Optimal decisions from very few samples. In *Proc. 31st annual CogSci.*

Yalcin, S. (2010). Probability Operators. *Phil. Compass*, *5*(11).