The effect of linking assumptions and number of response options on inferred scalar implicature rate

Masoud Jasbi[1], Brandon Waldon[1], & Judith Degen[1]

[1] Stanford University

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line. Enter author note here.

Correspondence concerning this article should be addressed to Masoud Jasbi, Postal address. E-mail: my@email.com

Abstract

Enter abstract here. Each new line herein must be indented, like this line.

Keywords: scalar implicature; methodology; linking assumption; experimental pragmatics; truth-value judgment task

Word count: X

The effect of linking assumptions and number of response options on inferred scalar implicature rate

Introduction

The past 15 years have seen the rise and development of a bustling and exciting new field at the intersection of linguistics, psychology, and philosophy: *experimental pragmatics* (Bott & Noveck, 2004; Breheny, Katsos, & Williams, 2006; Degen & Tanenhaus, 2015; Geurts & Pouscoulous, 2009; Grodner, Klein, Carbary, & Tanenhaus, 2010; Huang & Snedeker, 2009; I. A. Noveck & Reboul, 2008) **XXX ADD MORE**. Experimental pragmatics is devoted to experimentally testing theories of how language is used in context. How do listeners draw inferences about the – often underspecified – linguistic signal they receive from speakers? How do speakers choose between the many utterance alternatives they have at their disposal?

The most prominently studied phenomenon in experimental pragmatics is undoubtedly *scalar implicature*. Scalar implicatures arise in virtue of a speaker producing the weaker of two ordered scalemates (hornXXX; **???**, **???**; Grice, 1975). Examples are provided in (1) and (2).

1.

- *Utterance:* Some of her pets are cats.
- *Implicature:* Some, but not all, of her pets are cats.
- *Scale:*

2.

- *Utterance:* She owns a cat or a dog.
- *Implicature:* She owns a cat or a dog, but not both.
- *Scale:*

A listener, upon observing the utterances in (1a) and (2a), typically infers that the speaker intended to convey the meanings in (1b) and (2b), respectively. Since Grice (1975), the agreed-upon abstract rationalization the listener could give for their inference goes something like this: the speaker could have made a more informative statement by producing the stronger alternative (e.g., *All of her pets are cats.*). If the stronger alternative is true, they should have produced it to comply with the Cooperative Principle. They chose not to. I believe the speaker knows whether the stronger alternative is true. Hence, it must not be true.

Because the basic reconstruction of the inference is much more easily characterized for scalar implicatures than for other implicatures, scalar implicatures have served as a test bed for many questions in experimental pragmatics, including, but not limited to:

1. Are scalar inferences default inferences, in the sense that they arise unless blocked by (marked) contexts (Degen, 2015; Horn, 1984; Levinson, 2000)?

2. Are scalar inferences default inferences, in the sense that they are computed automatically in online processing and only cancelled by context in a second effortful step if required by context) [Bott and Noveck (2004);Breheny et al. (2006);Degen and Tanenhaus (2016);Grodner et al. (2010);Huang and Snedeker (2009);Politzer-Ahles and Fiorentino (2013);Tomlinson2013]?

3. What are the (linguistic and extra-linguistic) factors that affect whether a scalar implicature is derived [Zondervan (2010);Degen and Tanenhaus (2015); Degen and Tanenhaus (2016); Degen (2015); Degen and Goodman (2014); Bergen and Grodner (2012); Breheny et al. (2006); Breheny, Ferguson, and Katsos (2013);Marneffe and Tonhauser (2016);De Neys and Schaeken (2007);Bonnefon, Feeney, and Villejoubert (2009);Chemla2011;Potts2015]?

4. How much diversity is there across implicature types, and within scalar implicatures across scale types, in whether or not an implicature is computed (Doran, Ward, Larson, McNabb, & Baker, 2012; Tiel, Miltenburg, Zevakhina, & Geurts, 2014)?

5. At what age do children acquire the ability to compute implicatures (Barner, Brooks, & Bale, 2011; Katsos & Bishop, 2011; Frank; Musolino, 2004; Noveck, 2001; Papafragou & Tantalou, 2004)?

In addressing all of these questions, it has been crucial to obtain estimates of **implicature rates**. For 1., implicature rates from experimental tasks can be taken to inform whether scalar implicatures should be considered default inferences. For 2., processing measures on responses that indicate implicatures can be compared to processing measures on responses that indicate literal interpretations. For 3., contextual effects can be examined by comparing implicature rates across contexts. For 4., implicature rates can be compared across scales (or across implicature types). For 5., implicature rates can be compared across age groups.

A standard measure that has stood proxy for implicature rate across many studies is the proportion of "pragmatic" judgments in truth-value judgment paradigms [Bott and Noveck (2004);Noveck (2001);Noveck and Posada (2003);Chemla and Spector (2011);Geurts and Pouscoulous (2009);Degen and Tanenhaus (2015);De Neys and Schaeken (2007);Degen2014]. In these kinds of tasks, participants are provided a set of facts, either presented visually or via their own knowledge of the world. They are then asked to judge whether a sentence intended to describe those facts is true or false (or alternatively, whether it is right or wrong, or they are asked whether they agree or disagree with the sentence). The crucial condition for assessing implicature rates in these kinds of studies typically consists of a case where the facts are such that the stronger alternative is true and the target utterance is thus also true but underinformative. For instance, Bott and Noveck (2004) asked participants to judge sentences like "Some elephants are mammals", when world knowledge dictates that all elephants are mammals. Similarly, Degen and Tanenhaus (2015) asked participants to judge sentences like "You got some of the gumballs" in situations where the visual evidence indicated that the participant received all the gumballs from a gumball machine. In these kinds of scenarios, the story goes, if a participant responds "FALSE", that indicates that they computed a scalar implicature, eg to the effect of “Not all elephants are mammals” or “You didn’t get all of the gumballs”, which is (globally or contextually) false. If instead a participant responds “TRUE”, that is taken to indicate that they interpreted the utterance literally as ‘Some, and possibly all, elephants are mammals’ or “You got some, and possibly all, of the gumballs”.

Given the centrality of the theoretical notion of “implicature rate” to much of experimental pragmatics, there is to date a surprising lack of discussion of the basic assumption that it is adequately captured by the proportion of FALSE responses in truth-value judgment tasks (but see (**???**); Geurts and Pouscoulous (2009); Degen and Goodman (2014); Katsos and Bishop (2011)). Indeed, the scalar implicature acquisition literature was shaken up when Katsos and Bishop (2011) showed that simply by introducing an additional response option, children started looking much more pragmatic than had been previously observed in a binary judgment paradigm. (**???**) allowed children to distribute 1, 2, or 3 strawberries to a puppet depending on “how good the puppet said it”. The result was that children gave on average fewer strawberries to the puppet when he produced underinformative utterances compared to when he produced literally true and pragmatically felicitous utterances, suggesting that children do, in fact, display pragmatic ability even at ages when they had previously appeared not to.

But this raises an important question: in truth-value judgment task, how do we know whether an interpretation is literal or the result of an implicature computation? The binary choice task typically used is appealing in part because it allows for a direct mapping from response options – TRUE and FALSE – to interpretations – literal and pragmatic. That the seeming simplicity of this mapping is illusory becomes apparent once a third response option is introduced, as in the Katsos and Bishop (2011) case. How is the researcher to interpret the intermediate option? Katsos and Bishop (2011) grouped the intermediate option with the negative endpoint of the scale for the purpose of categorizing judgments as literal vs. pragmatic. But it seems just as plausible that they could have grouped it with the positive endpoint of the scale and taken the hard line that only truly FALSE responses constitute a full-fledged implicature. The point here is that there has been remarkably little consideration of **linking functions** between behavioral measures and theoretical constructs in experimental pragmatics, a problem in many subfields of psycholinguistics (**???**). We argue that it is time to engage more seriously with these issues.

We begin by reporting an experiment that addresses the following question: do the number of response options provided in a truth-value judgment task and the way that responses are grouped into pragmatic ("SI") and literal ("no SI") change inferences about scalar implicature rates? Note that this way of asking the question presupposes two things: first, that whatever participants are doing in a truth-value judgment task, the behavioral measure can be interpreted as providing a measure of **interpretation**. And second, that listeners either do or do not compute an implicature on any given occasion. In the Discussion we will discuss both of these issues. First, following Degen and Goodman (2014), we will offer some remarks on why truth-value judgment tasks are better thought of as measuring participants' estimates of speakers' **production** probabilities. This will suggest a completely different class of linking functions. And second, we discuss an alternative conception of scalar implicature as a probabilistic phenomeonen, a view that has recently rose to prominence in the subfield of probabilistic pragmatics. This alternative conception of scalar implicature, we argue, affords developing and testing quantitative linking functions in a rigorous and motivated way.

Consider a setup in which a listener is presented a card with a depiction of either one or two animals (see the figure below for an example). As in a standard truth-value judgment task, the listener then observes an underinformative utterance about this card (e.g., "There is a cat or a dog on the card") and is asked to provide a judgment on a scale from 2 to 5 response options, with endpoints "wrong" and "right". In the binary case, this reproduces the standard truth-value judgment task. **XXX say briefly sth about wrong/right vs true/false and agree/disagree**. The figure below exemplifies (some of) the researcher's options for grouping responses. Under what we will call the "Strong link" assumption, only the negative endpoint of the scale is interpreted as evidence for a scalar implicature having been computed. Under the "Weak link" assumption, in contrast, any response that does not correspond to the positive endpoint of the scale is interpreted as evidence for a scalar implicature having been computed. Intermediate grouping schemes are also possible, but these are the ones we will consider here. Note that for the binary case, the Weak and Strong link return the same categorization scheme, but for any number of response options greater than 2, the Weak and Strong link can in principle lead to differences in inferences about implicature rate.

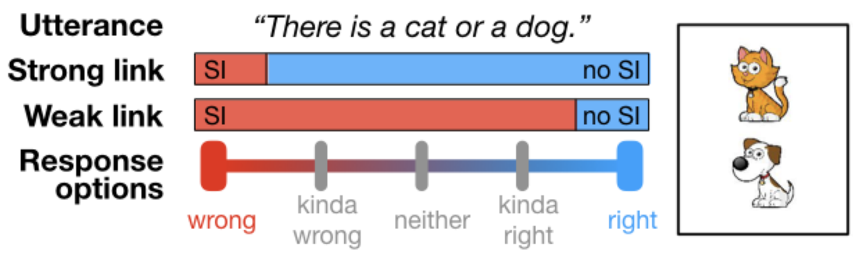


Figure 1. Strong and weak link from response options to researcher inference about scalar implicature rate, exemplified for the disjunctive utterance when the conjunction is true.

Let's examine an example. Assume three response options (wrong, neither, right). Assume further that a third of participants each gave each of the three responses, i.e., the distributions of responses is 1/3, 1/3, and 1/3. Under the Strong link, we infer that this task yielded an implicature rate of 2/3. Under the Weak link, we infer that this task yielded an implicature rate of 1/3. This is quite a drastic difference if we are for instance interested in whether scalar implicatures are inference defaults and we would like to interpret an implicature rate of above an arbitrary threshold (e.g., 50%) as evidence for such a claim. Under the Strong link, we would conclude that scalar implicatures are not defaults. Under the Weak link, we would conclude that they are.

In the experiment reported in the following section, we presented participants with exactly this setup. Different groups of participants were presented with different numbers of response options. We categorized their responses according to the Weak and the Strong link and tested whether number of response options and categorization scheme leads to different conclusions about implicature rates.

Experiment

Methods

Participants. 200 participants were recruited using Amazon Mechanical Turk (binary=50, ternary53, quaternary=43, quinary=54). No participant was excluded from the final analysis.

Procedure. The study was administered online through Amazon Mechanical Turk. Participants were introduced to a set of cards with pictures of one or two animals (Figure 2). They were told that a blindfolded fictional character called Bob is going to guess what animals are on the card. On each trial, participants saw a card as well as a sentence representing Bob's guess. For example, they saw a card with a cat on it and read the sentence "There is a cat on the card." The study ended after 24 trials. At the end participants optionally provided demographic information. We also asked participants if they had any prior training in logic. You can access and view on the study's online repository.

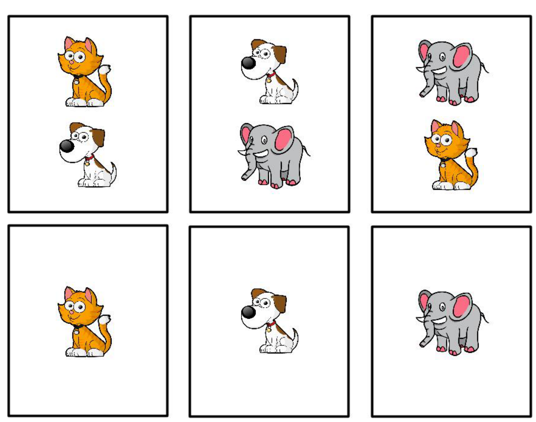


Figure 2. Cards used in the connective guessing game.

Design and Materials. The study had two main manipulaitons within participants: the type of card and the type of guess. There were two types of cards. Cards with only one animal on them and cards with two animals. Animals were chosen from the following set: cat, dog, and elephant There were three types of guesses: simple (e.g. *There is a cat*), conjunctive (e.g. *There is a cat and a dog*), and disjunctive (e.g. *There is a cat or a dog*). In each trial, the animal labels used in the guess and the animal images on the card may have no overlap (e.g. Image: cat, Guess: *There is an elephant*), a partial overlap (e.g. Image: cat, Guess: *There is a cat or a dog*), or a total overlap (e.g. Image: cat and dog, Guess: *There is a cat or a dog*). Crossing the number of animals on the card, the type of guess, and the overlap between the guess and the card results in 12 different possible trial types. We chose 8 trial types (Figure 3), balancing the number of one-animal vs. two-animal cards, simple vs. connective guesses, and expected true vs. false trials. Three trials were randomly selected from each of the 8 trial-types, for a total of 24 trials. The order of these 24 trials was randomized as well.

Participants could derive implicatures in two trial types. First, the trial type in which two animals were present on the card (e.g. cat and dog) but Bob guessed only one of them (e.g. "there is a cat"). Such trials can have a literal interpretation (cat and possibly more) or an exhaustive interpretation (only cat). We refer to them as "exhaustive". The second trial type with implicatures was the one in which two animals were on the card (e.g. cat and dog) and Bob used a disjunciton (e.g. cat or dog). These trials can have a literal (inclusive) interpretation (e.g. cat or dog or both), or an exclusive interpretation (e.g. cat or dog, not both). We refer to these trials as "scalar". The following four trial types were used as experimental control: two trial types where there was no overlap between the guess (e.g. elephant) and the animal(s) on the card (e.g. cat, cat and dog); and two trial types where the animal(s) on the card were correctly guessed. For example, if there was only a cat on the card, Bob said "there is a cat" and if there was a cat and a dog, Bob said "there is a cat and a dog". Since the fictional character was blindfolded and did not see the outcome of the game, the ignorance inference commonly associated with disjunction was already common ground in the experimental setting. If the character was seeing the cards or knew what was on them, a disjunction would have violated the expectation that the speaker does not know which alternative actually holds. Our study controls for the possibile effect of ignorance violations on exclusivity and exhaustive inferences.

The study also had a between participant manipulation of the number of response options in the forced choice task. Participants were randomly assigned to one of four different conditions. The conditions differed with respect to the number of response options: binary (wrong vs. right), ternary (wrong, neither, right), quaternary (wrong, kinda wrong, kinda right, right), and quinary (wrong, kinda wrong, neither, kinda right, right). We wanted to see if the number of response options in the forced choice task would affect our estimate of the task's "implicature rate".

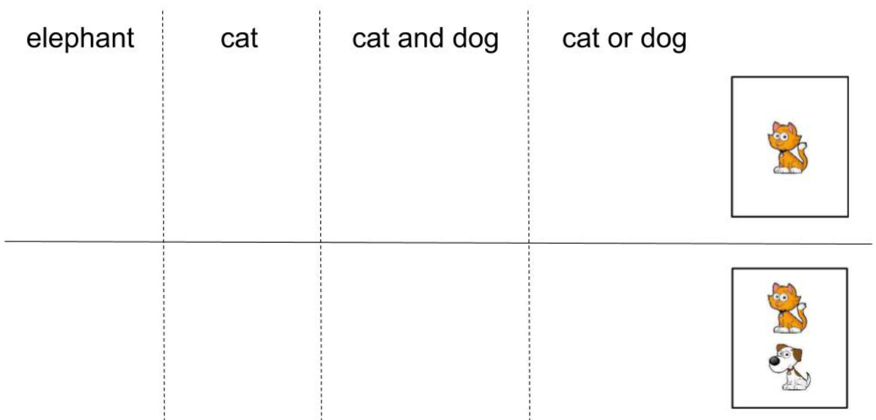


Figure 3. Trial types represented by example cards and guesses.

Results

Here we present the proportion that participants chose the response options in each of the 8 trial types of the binary, ternary, quaternary, and quinary tasks. Figure 4 shows the proportion of "right" and "wrong" responses in the binary task. Starting from the leftmost column, participants considered a guess "wrong" if the guessed animal was not on the card.

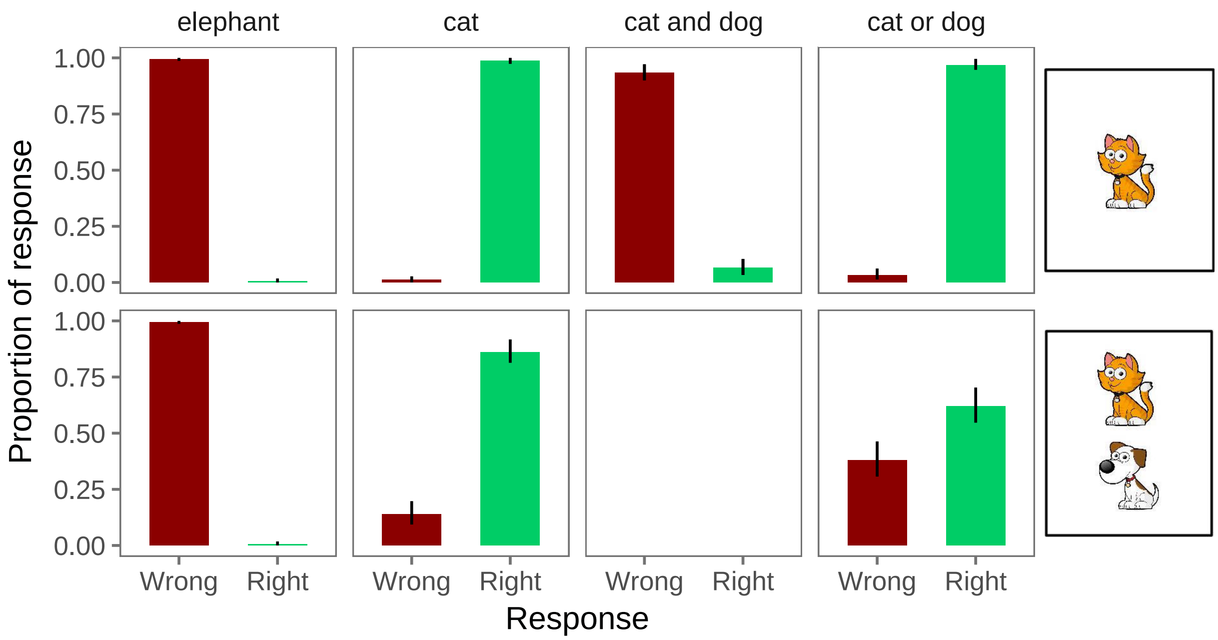


Figure 4. Proportion responses for the two-alternative (binary) forced choice judgments in the guessing game.

Moving to the second column, participants considered a guess "right" if the animal on the card was mentioned. However, if only one of the two animals on the card was mentioned (exhaustive trials), 14% of the times participants considered the guess "wrong". Moving to the third column, if a conjunction of animals was guessed while only one animal was on the card, participants considered the guess to be "wrong". If a conjunction of animals was guessed and both animals were present on the card, all participants considered the guess to be "right" as expected. Moving to the forth column, if a disjunction of animals was guessed and only one of the animals was on the card, participants considered the guess to be "right" almost all the time. However, if both animals were present (scalar trials), 38% of the times participants considered the guess to be "wrong".

Figure 5 shows the proportion of "right", "neither", and "wrong" responses in the ternary task. Similar to the binary task, participants considered a guess wrong when the mentioned animal was not on the card. They considered the guess "right" when the mentioned animal was on the card. However, in exhaustive trials when the fictional character

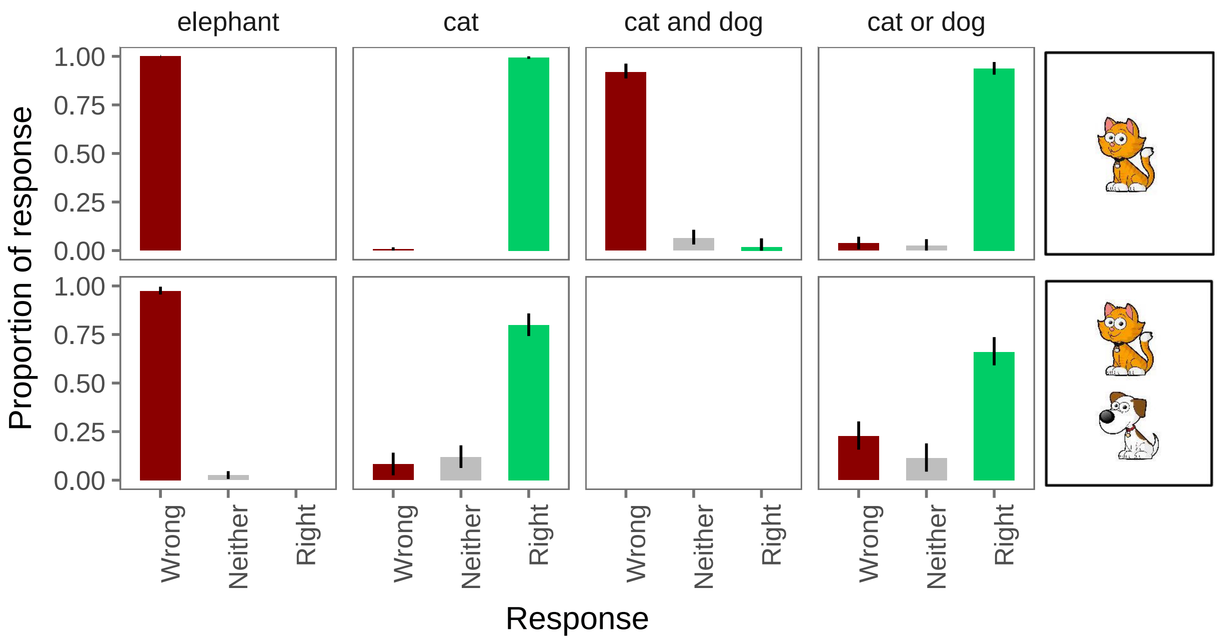


Figure 5. Proportion responses for the three-alternative (ternary) forced choice judgments in the guessing game.

only guessed one of the two animals on the card, participants considered the guess "wrong" 8% of the time and neither wrong nor right 12% of the time. If a conjunction of animals was guessed and only one animal was present on the card, participants considered the guess "wrong". As expected, when a conjunction was used and both animals were present, participants considered the guess "right". Similarly, participants considerd the guess "right" when a disjunction was used and only one of the animals was on the card. However, in scalar trials that both animals were on the card and a disjunction was gueassed, participants judged the guess "wrong" 23% of the time and "neither" 11% of the time.

Figure 6 shows the proportion of "right", "kinda right", "kinda wrong", and "wrong" responses in the quaternary task. Similar to the results seen previously, the control trials turned outas expected. Participants considered a guess "wrong" if the animal guessed was not on the card and "right" if it was the only animal on the card. If a conjunctin of animals was guessed and both animals were on the card the guess was "right". However, when only one of the animals on the card was guessed (exhaustive trials), participants judged the guess

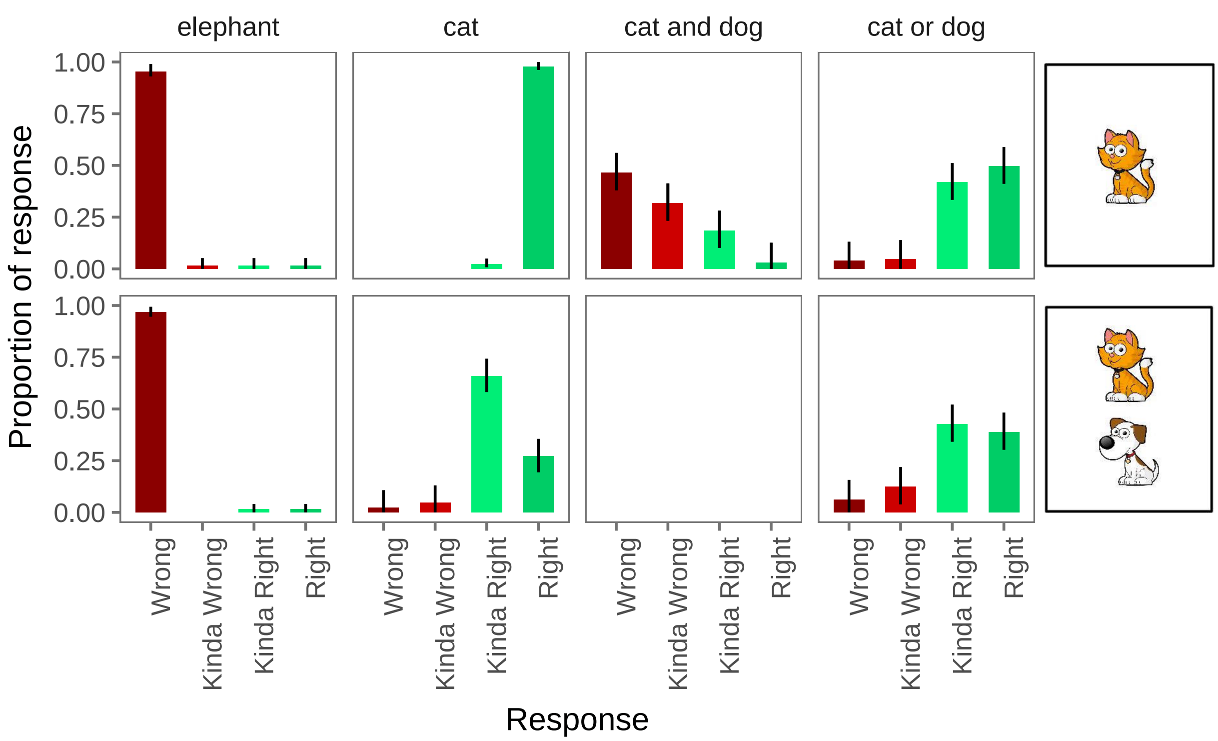


Figure 6. Proportion responses for the four-alternative (quatenary) forced choice judgments in the guessing game.

"wrong" 2% of the time, "kinda wrong" 5% of the time, and "kinda right" 66% of the times. Perhaps surprisingly, when a conjunction was used and only one of the animals was on the card, participants considered the guess "wrong" most of the time, but they often considered it "kinda wrong" or even "kinda right". This suggests that perhaps participants considered a notion of partially true or correct statement in our experimental setting. Disjunctive guesses with one or two animals on the card showed similar response patterns with participants choosing the "kinda right" and "right" options most of the time. When both animals were on the card with a disjunctive guess (scalar trials), participants judged the guess "wrong" 6% of the time, "kinda wrong" 12% of the time, and "kinda right" 43% of the times.

Finally, Figure 7 shows the proportion of "right", "kinda right", "neither", "kinda wrong", and "wrong" responses in the quinary task. Since the results for the control trials were identical to previous tasks, we do not repeat them here. In exhaustive trials where two

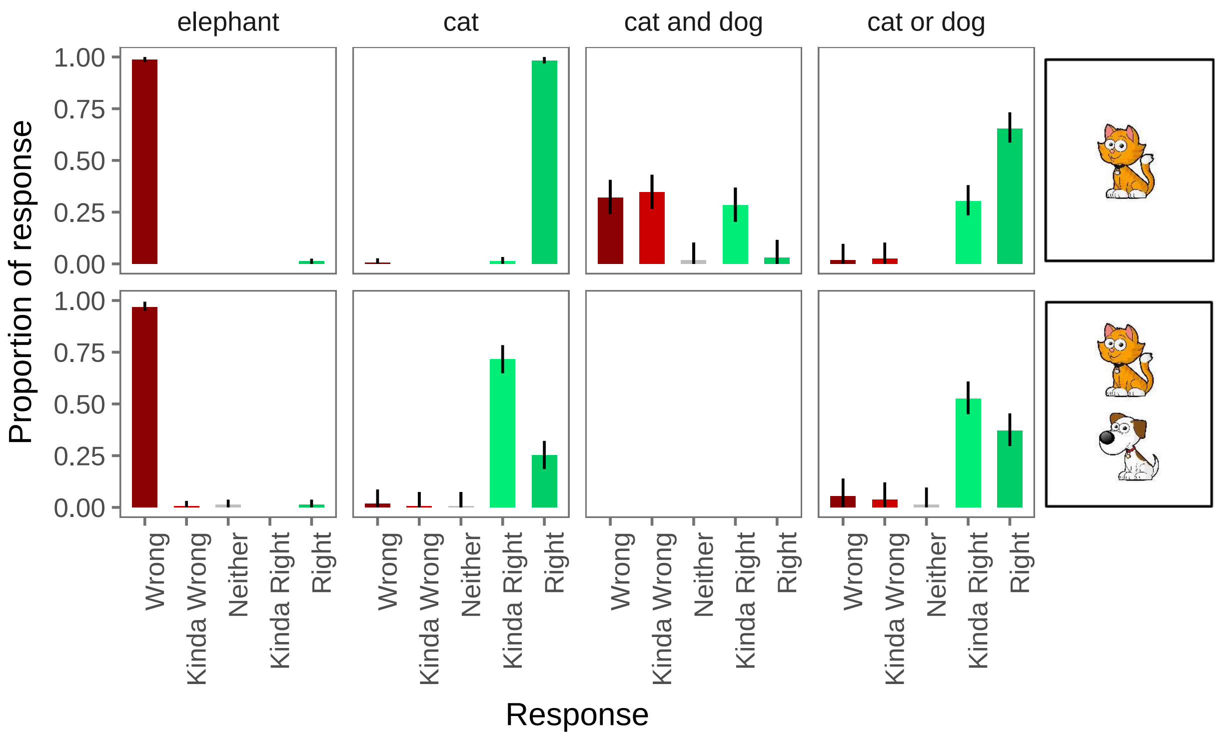


Figure 7. Proportion responses for the five-alternative (quinary) forced choice judgments in the guessing game.

animals were on the card and only one of them was guessed, participants chose "kinda right" the majority of times (72%). Again perhaps surprisingly, when only one animal was on the card and the guess was a conjunction, responses were equally split among "wrong", "kinda wrong", and "kinda right" responses. With disjunctive guesses, partitipants were more likely to choose "right" and "kinda right" options. When only one animal was on the card, participants considered the disjunctive guess as "right" more often. When both animals were on the card (scalar trials), participants judged the disjunctive guess as "kinda right" 52% of the time.

Comparing the response patterns in binary to quinary tasks (Figures 4 to 7), we can observe that in implicature trials, participants are less likely to choose the endpoints of the scale (i.e. "wrong" and "right") as they are given more intermediate options. This raises the possibilty that the strict "wrong/right" responses in the tasks with fewer options (binary and ternary) were not accurately reflecting participant judgments of implicature trials and pushed their judgments to the extreme ends. More generally, the data suggest that participant judgments in contexts that give rise to implicatures fall within the intermrediate level of the response scale and forced choice tasks with few options (especially binary tasks) may risk depicting judgments more extreme than they actually are.

Analysis

Our primary goal in this study was to check whether the estimated “implicature rate” in the experimental task is affected by the linking assumption and the number of response options available in the task. Our analysis in this section focuses on these three elements. As mentioned before, two trial types were predicted to include pragmatic implicatures. First, trials where two animals were on the card but the fictional character guessed with a disjunction (scalar); for example “cat or dog” when the card has both a cat and a dog on it. Second, trials where there were two animals on the card but the character guessed only one (exhaustive); for example “cat” when the card had a cat and a dog on it. We called such trials “exhaustive”. In our assessment of implicature rate, we focus on these two trial types.

We considered two linking assumptions. First we defined a weak (lenient) linking assumption in which any response lower than the maximum point on the scale (i.e. “right”) is considered evidence for implicature computation. Second, we defined a strong (strict) linking assumption that only considered the lowest point on the scale (i.e. “wrong”) as evidence for implicature computation. For each condition in our study (binary, ternary, quaternary, and quinary) and each implicature trial type (exhaustive and scalar), we computed a weak and a strong implicature rate. Figure 8 shows these computed rates.

Comparing the strong and weak rows on Figure 8, we see that a weak linking assumption tends to estimate higher implicature rates, especially in tasks with more response options. With a strong linking assumption, the binary and possibly ternary judgment tasks derive higher implicature rates than quaternary and quinary tasks. With a weak linking assumption, the pattern is reversed. Quaternary and quinary tasks estimate higher rates than binary and ternary tasks. The patterns show that estimates of "implicature rate" depend on linking assumptions and the number of responses available to participants in the study.

Comparing the exhaustive and scalar columns of Figure 8, we see that with a strong linking assumption, there are slightly higher rates for scalar implicatures in the binary and ternary tasks. With a weak linking assumption, there may be slightly higher rates for scalar implicatures in the binary and ternary while the rates may be lower in the quaternary and quinary tasks. In what follows, we test the effect of linking assumption and response options on exhaustive and scalar implicature rates more formally.

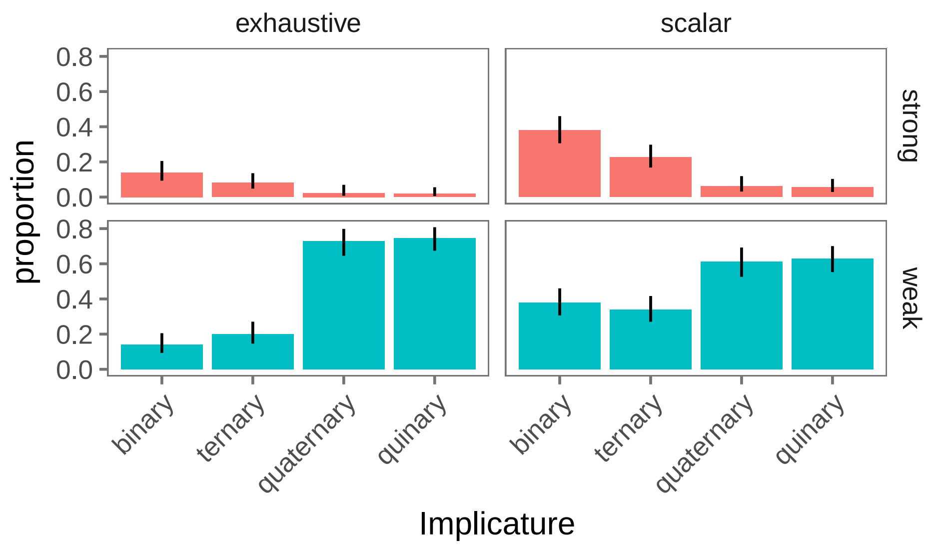


Figure 8. Implicature rate in exhaustive and scalar trials of the binary, ternary, quaternary, and quinary versions of the guessing game, computed once with a strong linking assumption and once with a weak linking assumption.

For our formal analysis, we use a binomial mixed effects model with the fixed effect of response type (binary, ternary, quaternary, quinary), linking assumption (strong vs. weak), and trial type (exhaustive vs. scalar), as wellas the random intercept for participants as well as random intercept and slope for items (the card they saw).

All the analytical decisions described here were pre-registered and you can access them through the following link: https://aspredicted.org/tq3sz.pdf

- Was logical training a significant predictor of "implicature rate"?

 # Discussion

On the traditional view of the link between implicature and behavior in sentence verification tasks, scalar implicature is conceptualized as a binary, categorical affair - that is, an implicature is either "calculated" or it isn't, and the behavioral reflexes of this categorical interpretation process should be straightforwardly observed in experimental paradigms. This assumption has concerning implications for how we must approach analysis of variation in behavior on a truth value judgment task; for example, why did the majority of respondents in the binary condition of our experiment answer "Right" to an utterance of cat or dog when the card had both a cat and a dog on it?

To explain the data on the traditional view, are forced to say that a) not all participants calculated the implicature; or that b) some participants who calculated the implicature did not choose the anticipated response (i.e. "Wrong") due to some other cognitive reflex which "overrode" the implicature; or some mixture of (b) and (c). We might similarly posit that one or both of these factors underlie the variation in the ternary, quatenary, and quinary conditions (e.g. why were participants roughly split between "Right" and "Kind of right" when the utterance was cat or dog and the card had a cat and a dog?). However, the best we can hope for on this approach is an analysis which traces the general qualitative patterns in the data.

We contrast the above view of implicature and its behavioral reflexes with an alternative linking hypothesis which assumes that participants' behavior can be represented using the model of a soft-optimal pragmatic speaker in the RSA framework. This alternative linking hypothesis contrasts with the traditional view in it is rooted in a quantitative formalization of pragmatic competence which provides us a continuous measure of pragmatic reasoning. Recall that in RSA, pragmatically competent listeners are modeled as a continuous probabilistic distribution of possible meanings given an utterance which that listener hears. The probability with which this listener L_1 ascribes a meaning s to an utterance u depends upon a prior probability distribution of potential states of the world P_w, and upon reasoning about the communicative behavior of a speaker S_1. S_1 in turn is modeled as a continuous probabilistic distribution of possible utterances given an intended state of affairs the speaker intends to communicate. This distribution is sensitive to a rationality parameter α, the production cost C of potential utterances, and a representation of a literal listener L_0 whose interpretation of an utterance is in turn a function of that utterance's truth conditional content $[[u]](s)$ and her prior beliefs about the state of the world $P_w(s)$.

```
$P_{L_1}(s | u) \propto P_{S_1}(u | s) * P_w(s)$

$P_{S_1}(u | s) \propto exp(\alpha(log(L_0(s | u)) - C(u))) $

$P_{L_0}(s | u) \propto [[u]](s) * P_w(s)$
```

In this framework, individuals never categorically draw (or fail to draw) pragmatic inferences about the utterances they hear. For example, exclusivity readings of disjunction or are represented in RSA as relatively low conditional probability of a conjunctive meaning on the P_L distribution, given an utterance of or. Thus, it is not even possible to talk about "rate" of implicature calculation in the RSA framework. The upshot, as we show below, is that this view of pragmatic competence does allow us to talk explicitly and quantitatively about rates of observed behavior in sentence verification tasks.

First, following Degen & Goodman (2014), we proceed on the assumption that behavior on sentence verification tasks, such as truth value judgment tasks, is best modeled as a function of an individual's mental representation of a cooperative interlocutor (S_1 in the language of RSA) rather than of a pragmatic listener who interprets utterances (P_{L_1}). In their paper, Degen & Goodman argue that sentence verification tasks are relatively more sensitive to contextual manipulations (such as manipulation of the Question Under Discussion) than are sentence interpretation tasks, and that this follows if sentence interpretation tasks - but not sentence verification tasks - require an additional layer of counterfactual reasoning about the intentions of a cooperative speaker.

A main desideratum of a behavioral linking hypothesis given the RSA view of pragmatic competence is to transform continuous probability distributions into categorical outputs (e.g. responses of "Right"/"Wrong" in the case of the binary condition of our experiment). For a given utterance u and an intended communicated meaning w, S_1(u | w) outputs a conditional probability of u given w. For example, in the binary condition of our experiment where a participant evaluated cat or dog when there were both animals on the card, the participant has access to the mental representation of S_1 and hence to the S_1 conditional probability of hearing the utterance cat or dog given a dog and cat card: S_1(cat or dog | cat and dog). According to the linking hypothesis advanced here, the participant provides a particular response to u if the RSA speaker probability of u lies within a particular probability interval, given an observed state of the world (i.e. the configuration of animals on the card in our experiment). We model a responder, R, who in the binary condition responds "Right" to an utterance u in world w just in case S_1(u | w) exceeds some probability threshold θ:

R(u, w, θ)

= "Right" iff S_1(u | w) $> \theta$

= "Wrong" otherwise

In the experiment conditions where there are more than two choices, we model the range of possible behavioral responses for R with the introduction of intermediate probability thresholds. For example, in the ternary condition, R(u, w, θ_1 , θ_2) is "Right" iff S_1(u | w) $>$ θ_1 and "Neither" iff $\theta_1 > S_1$(u | w) $> \theta_2$. To fully generalize the model to our five experimental conditions, we say that R takes as its input an utterance u, a world state w, and a number of threshold variables dependent on a variable c, corresponding to the experimental condition in which the participant finds herself (e.g. the range of possible responses available to R).

Given c = "ternary"

R(u, w, θ_1 , θ_2)

= "Right" iff S_1(u | w) > θ_1

= "Neither" iff $\theta_1 > S_1$(u | w) > θ_2

= "Wrong" otherwise

Given c = "quatenary"

R(u, w, θ_1 , θ_2, θ_3)

= "Right" iff S_1(u | w) > θ_1

= "Kinda Right" iff $\theta_1 > S_1$(u | w) > θ_2

= "Kinda Wrong" iff $\theta_2 > S_1$(u | w) > θ_3

= "Wrong" otherwise

Given c = "quinary"

R(u, w, θ_1 , θ_2, θ_3. θ_4)

= "Right" iff S_1(u | w) > θ_1

="Kinda Right" iff $\theta_1 > S_1$(u | w) > θ_2

= "Neither" iff $\theta_2 > S_1$(u | w) > θ_3

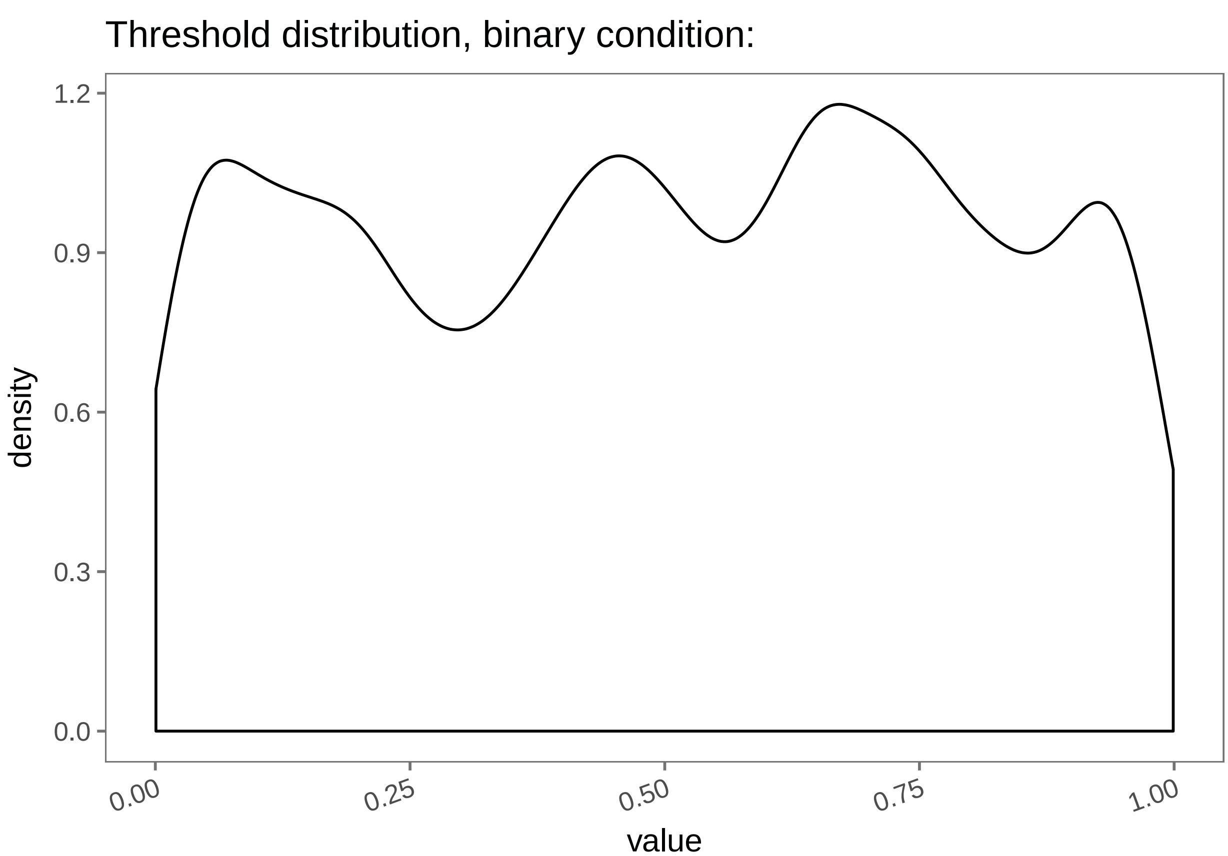
= "Kinda Wrong" iff $\theta_3 > S_1$(u | w) > θ_4

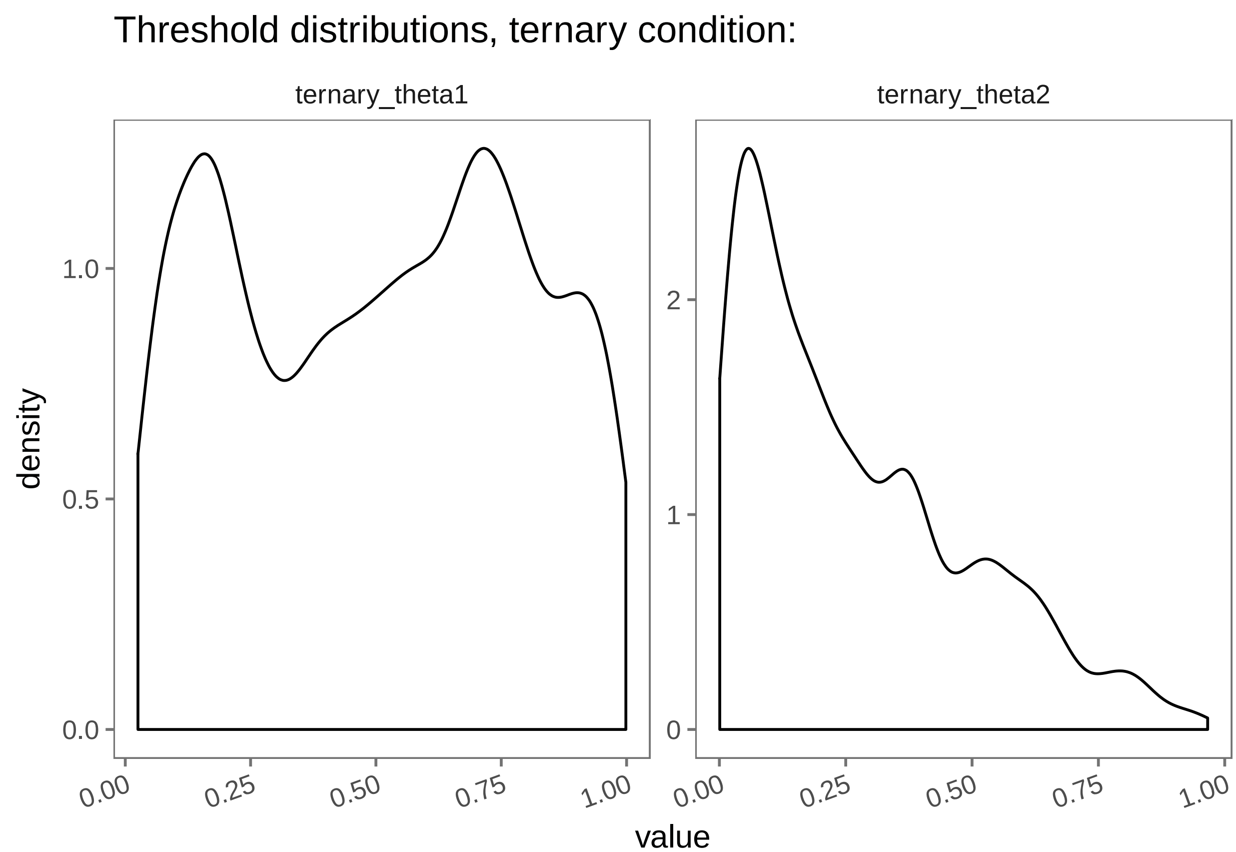
= "Wrong" otherwise

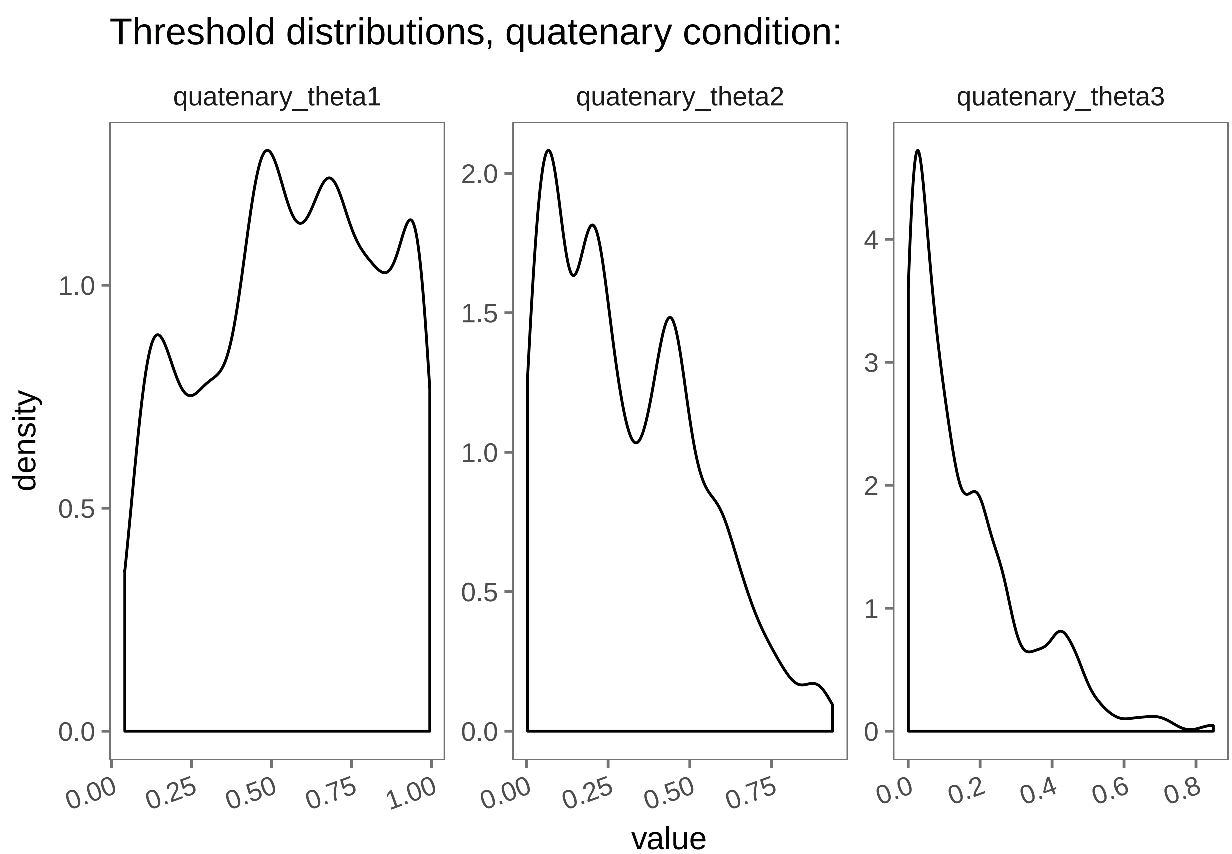
Bayesian statistical methods provide us a means for estimating the values of these probability thresholds in our RSA model. The basis for the model is a set of possible states of the world, given a universe of three animals - X, Y, and Z - that each may be on some card. We next define a set of possible sentences a speaker might utter, assuming the speaker intends to communicate which animals are on the card. We assume a uniform prior probability of different states of the world and a uniform cost function on utterances. We define a literal listener L_0, a pragmatic speaker S_1, and a responder R according to our definitions above. Lastly, and assuming a uniform prior distribution over possible values of probability thresholds, we use Bayesian inference to recover a posterior distribution of these thresholds in each experimental condition, given the actual observed rate of response in each condition of the experiment. The results of this parameter estimation analysis are shown in the figures below, where the X axis of each figure corresponds to a threshold value between 0 and 1 and the Y axis corresponds to the posterior probability density of possible values of the threshold.

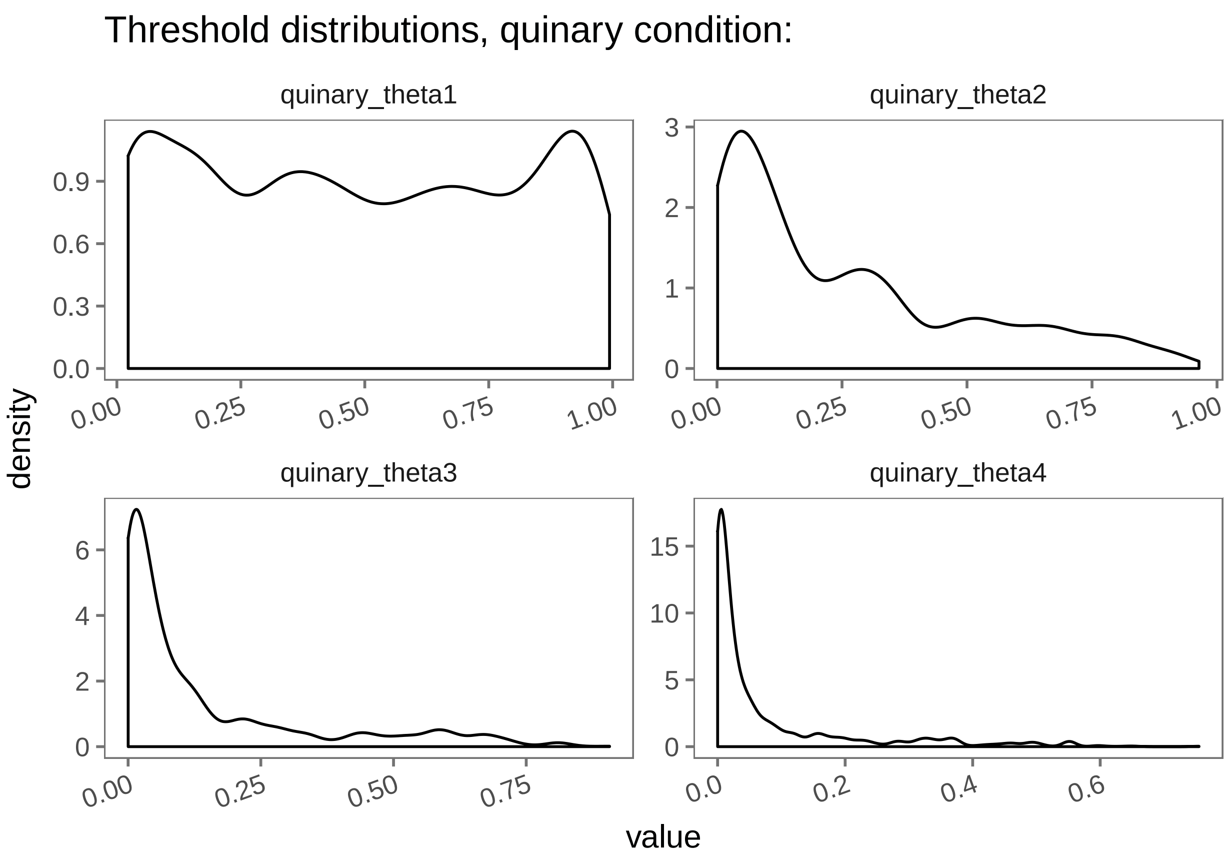
```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

```
## Warning in `==.default`(Parameter, c("quatenary_theta1",
## "quatenary_theta2", : longer object length is not a multiple of shorter
## object length
```




Threshold distributions, ternary condition:
ternary_theta1
ternary_theta2
density
value
0.0
0.5
1.0
0
1
2
0.00
0.25
0.50
0.75
1.00


Threshold distributions, quatenary condition:
quatenary_theta1
quatenary_theta2
quatenary_theta3
density
value
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
0.8



The above analysis is a proof of concept for the following idea: by relaxing the assumptions of the traditional view of scalar implicature (namely, that scalar implicatures either are or are not calculated, and that behavior on sentence verification tasks directly reflects this binary interpretation process), we can propose quantitative models of the variation in behavior we observe in experimental settings.

References

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84–93. doi:10.1016/j.cognition.2010.10.010

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology. Learning, Memory, and*

Cognition, *38*(5), 1450–60. doi:10.1037/a0027850

Bonnefon, J.-F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: scalar inferences in face-threatening contexts. *Cognition*, *112*(2), 249–58. doi:10.1016/j.cognition.2009.05.005

Bott, L., & Noveck, I. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*(3), 437–457. doi:10.1016/j.jml.2004.05.006

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition*, *126*(3), 423–40. doi:10.1016/j.cognition.2012.11.012

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*(3), 434–63. doi:10.1016/j.cognition.2005.07.003

Chemla, E., & Spector, B. (2011). Experimental Evidence for Embedded Scalar Implicatures. *Journal of Semantics*, *28*(3), 359–400.

De Neys, W., & Schaeken, W. (2007). When People Are More Logical Under Cognitive Load - Dual Task Impact on Scalar Implicature. *Experimental Psychology*, *54*(2), 128–133. doi:10.1027/1618-3169.54.2.128

Degen, J. (2015). Investigating the distribution of 'some' (but not 'all') implicatures using corpora and web-based methods. *Semantics and Pragmatics*, *8*(11), 1–55. doi:10.3765/sp.8.11

Degen, J., & Goodman, N. D. (2014). Lost your marbles? The puzzle of dependent measures in experimental pragmatics. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 397–402.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature A constraint-based approach. *Cognitive Science*, *39*(4), 667–710. doi:10.1111/cogs.12171

Degen, J., & Tanenhaus, M. K. (2016). Availability of Alternatives and the Processing of

Scalar Implicatures: A Visual World Eye-Tracking Study. *Cognitive Science*, *40*(1), 172–201. doi:10.1111/cogs.12227

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, *88*, 124–154.

Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics and Pragmatics*, *2*, 1–34. doi:10.3765/sp.2.4

Grice,

H. P. (1975). Logic and Conversation. *Syntax and Semantics*, *3*, 41–58. Retrieved from http://books.google.com/books?hl=en{\&}lr={\&}id=hQCzOmaGeVYC{\&}oi=fnd{\&}pg=PA1

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*(1), 42–55. doi:10.1016/j.cognition.2010.03.014

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11–42). Washington: Georgetown University Press.

Huang, Y. T., & Snedeker, J. (2009). On-line interpretationf of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*, 376–415.

Katsos, N., & Bishop, D. V. M. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81. doi:10.1016/j.cognition.2011.02.015

Levinson, S. C. (2000). *Presumptive Meanings - The Theory of Generalized Conversational Implicature*. MIT Press.

Marneffe, M.-C. de, & Tonhauser, J. (2016). Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour. In E. Onea, M. Zimmermann, & K. von Heusinger (Eds.), *Questions in discourse*. Leiden: Brill Publishing.

Musolino, J. (2004). The semantics and acquisition of number words: integrating linguistic and developmental perspectives. *Cognition*, *93*(1), 1–41. doi:10.1016/j.cognition.2003.10.002

Noveck, I. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11074249

Noveck, I. A., & Reboul, A. (2008). Experimental pragmatics: a Gricean turn in the study of language. *Trends in Cognitive Sciences*, *12*(11), 425–431. doi:10.1016/j.tics.2008.07.009

Noveck, I., & Posada, A. (2003). Characterizing the Time Course of an Implicature: an Evoked Potentials Study. *Brain and Language*, *85*(2), 203–210. doi:10.1016/S0093-934X(03)00053-1

Papafragou, A., & Tantalou, N. (2004). Children's Computation of Implicatures. *Language Acquisition*, *12*(1), 71–82.

Politzer-Ahles, S., & Fiorentino, R. (2013). The Realization of Scalar Inferences: Context Sensitivity without Processing Cost. *PLoS ONE*, *8*(5). doi:10.1371/journal.pone.0063943

Tiel, B. van, Miltenburg, E. van, Zevakhina, N., & Geurts, B. (2014). Scalar diversity. *Journal of Semantics*. doi:10.1093/jos/ffu017

Zondervan, A. (2010). *Scalar implicatures or focus: an experimental approach* (PhD thesis). Universiteit Utrecht, Amsterdam.