

# Non-sinking marbles are wonky: defeasible world knowledge in language interpretation

Judith Degen (jdegen@stanford.edu)

Michael H. Tessler (mtessler@stanford.edu)

Noah D. Goodman (ngoodman@stanford.edu)

Department of Psychology, 450 Serra Mall  
Stanford, CA 94305 USA

## Abstract

World knowledge enters pragmatic utterance interpretation in complex ways. Sometimes, a speaker’s utterance suggests that listeners should disregard their world knowledge, yet current models of pragmatic interpretation either disregard the role of world knowledge or overestimate its role in interpretation. Here we provide an extension to the Rational Speech Act model of scalar implicature that captures whether listeners believe they are in an abnormal—or ‘wonky’—world after observing a speaker’s utterance, in which case they downweight their prior beliefs in the computation of speaker meaning. We show in four experiments that a) listeners have varying prior beliefs about the probability of various objects exhibiting an effect (e.g., marbles sinking), b) these beliefs influence listeners’ expectations about how many objects will show the effect after observing an utterance (like *Some of the marbles sank*), c) these beliefs influence scalar implicature strength, and d) listeners’ world wonkiness judgments are affected by the surprisal of the observed utterance under their prior beliefs. The extended model is the first quantitative model that accounts for how rational listeners should integrate world knowledge in pragmatic utterance interpretation, and provides a close match to the empirically obtained data.

**Keywords:** scalar implicature; world knowledge; prior beliefs; experimental pragmatics; computational pragmatics

How often do you think marbles would sink in water? Probably quite often, if not always. Now imagine reading *Max threw his favorite marbles in the water. Some of them sank*. Have you begun to reconsider your assumptions? Perhaps you now suspect that these marbles are in fact made of plastic or the water is enriched with something that makes them float? That is, that they are not just normal marbles in normal water. Here we explore how prior world knowledge enters into pragmatic utterance interpretation, and how this world knowledge is defeasible: some utterances lead listeners to conclude that the world under discussion is abnormal under their prior beliefs. We refer to such an abnormal world as a *wonky* world.

## Models of utterance interpretation

Recent Bayesian Rational Speech Act (RSA) (?, ?) and game-theoretic (?, ?) models that treat communication as a signaling game (?, ?) between a speaker and a listener have successfully captured listeners’ quantitative behavior on a number of pragmatic inference tasks, including ad hoc Quantity implicature [ref](#), M-implicature [ref](#), scalar implicature [ref](#), and embedded scalar implicatures [ref](#). In these models, the listener iteratively reasons about likely utterances a speaker will produce

who is trying to be informative with respect to a naïve listener. These models make clear predictions about how prior beliefs about states of the world should be integrated with listeners’ expectations about utterances a speaker is likely to produce to communicate a particular state of the world.

For concreteness, assume that  $S = \{s_0, s_1, s_2, \dots, s_{15}\}$  is the set of states of the world, where the subscript indicates the number of objects (e.g., marbles) that exhibit a certain effect (e.g., sinking). Assume further that  $U = \{u_{\text{all}}, u_{\text{none}}, u_{\text{some}}\}$ , the set of utterances *All/None/Some of the marbles sank*. The speaker chooses an utterance to convey  $s$  proportional to the soft-max expected utility of producing  $u$  to communicate  $s$ :

$$P_{\text{speaker}}(u|s) \propto \exp(\lambda \mathbb{E}(\ln P_{\text{lex}}(s|u))) \quad (1)$$

where  $P_{\text{lex}}(s|u)$  is the literal interpretation probability resulting from each utterance’s truth-functional meaning:  $F_u : s \mapsto \{0, 1\}$ . The listener’s task is to infer a distribution over  $S$ , given an utterance  $u$  produced by the above defined informative speaker. By Bayes’ rule:

$$P_{\text{listener}}(s|u) \propto P_{\text{speaker}}(u|s) \cdot P(s) \quad (2)$$

That is, the inferred listener probabilities are proportional to the product of both the speaker’s utterance probabilities and the listener’s prior beliefs in different numbers of marbles sinking. Using a uniform prior over the state space, this model has been very successful at capturing *scalar implicatures* (?, ?). These are inferences that arise in cases of utterances like *Some of the marbles sank*, which typically give rise to the inference that not all of the marbles sank. Scalar implicatures fall out of the fact that the speaker probability of producing  $u_{\text{some}}$  in  $s_{15}$  is low (because there is an alternative utterance  $u_{\text{all}}$  which has a higher probability conveying  $s_{15}$  because it is more informative about the state). However, the role of prior beliefs in RSA models remains under-explored, both with respect to scalar implicature computation as well as with respect to utterance interpretation more generally.

One important consequence of the standard RSA model is that the listener’s prior beliefs will not affect the interpretation of  $u$  very strongly, if at all, where the semantics of  $u$  is highly constraining, that is, where  $P_{\text{lex}}(s|u)$  is very peaked. For example, an utterance of *All of the marbles sank* is highly constraining because all of the probability mass ends up on  $s_{15}$  by virtue of the semantics of  $u_{\text{all}}$  alone, regardless of how

likely the listener believes marbles sink a priori.<sup>1</sup> In contrast, where the semantics of  $u$  is weak – that is, where  $P_{\text{lex}}(s|u)$  is less peaked – prior beliefs are predicted to have a large effect on the resulting listener belief distribution. For example, an utterance of *Some of the marbles sank*, produced in a situation in which any of 0 - 15 of 15 contextually established marbles could have sunk, semantically only restricts the state space by one state (that in which 0 marbles sank). In this case, listeners’ prior beliefs about sinking marbles will have a large effect on their posterior belief distribution. If the listener believes that marbles rarely sink, the utterance will be interpreted as conveying that fewer marbles sank than if the listener believes marbles almost always sink. The predictions this model makes for the interpretation of  $u_{\text{some}}$  – both for  $P_{\text{listener}}(s_{15}|u_{\text{some}})$  and for the expected value of  $P_{\text{listener}}(s|u_{\text{some}})$  as a function of  $P(s_{15})$  and the expected value of  $P(s)$ , respectively – are shown in Figure ?? . RSA predicts that the probability of the state in which all objects exhibit a certain effect increases with increasing  $P(s_{15})$ , such that for  $P(s_{15})$  close to 1,  $P_{\text{listener}}(s_{15}|u_{\text{some}})$  approaches 1. Relatedly, with increasing expected value of the prior belief distribution  $P(s)$ , so is the expected value of the posterior belief distribution predicted to increase, approaching 1).

However, intuition is at odds with this prediction: for example, ? (?) has observed that for events with very high prior probability of occurrence (e.g. marbles sinking), observing an utterance of *Some of the marbles sank* leads to very strong implicatures, that is, the subjective probability that not all of the marbles sank is intuitively close to 0. Given the previous success of RSA models, this constitutes a striking puzzle, and one we set out to solve here. In doing so, we pursue the intuition raised at the very beginning of this paper: that sometimes, the speaker’s utterance will lead the listener to infer that the world under discussion is wonky and she should therefore down-weight her prior beliefs in the computation of speaker meaning.

Our contribution is two-fold: first, we collect empirical estimates of  $P(s)$  and  $P_{\text{listener}}(s|u)$  to investigate the empirical effect of listeners’ prior beliefs on interpretation. Second, we extend the RSA model to incorporate a free variable  $\theta_{\text{wonky}}$  that captures the extent to which the listener believes the described event is wonky and she should thus discount her prior beliefs when interpreting  $u$ . We refer to this model as *wonky RSA* (*wRSA*) in contrast to *regular RSA* (*rRSA*). Wonkiness inferences in wRSA are triggered by the surprisal of a produced utterance  $u$ , given listeners’ prior beliefs, capturing that listeners expect speakers’ utterances to be both truthful and informative with respect to prior beliefs. To the extent that they are not, listeners will have to either infer that the speaker is being uncooperative, or else assume that they may need to revise their beliefs about the world. Here we pursue the latter possibility.

This paper is structured as follows. We first report the re-

<sup>1</sup>We ignore hyperbolic interpretations of universal quantifiers here.

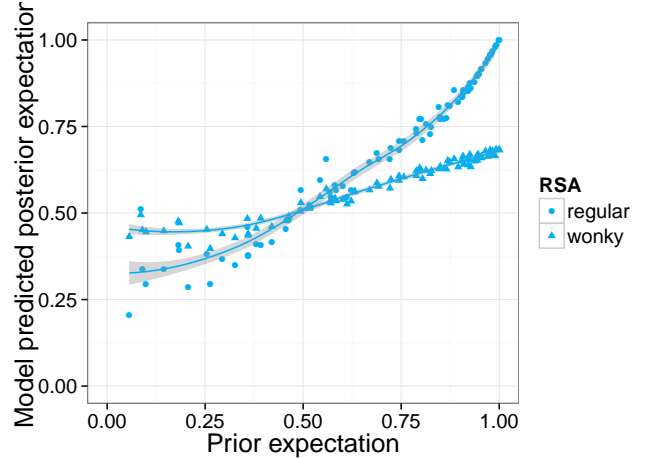


Figure 1: For each item, rRSA and wRSA model predicted mean empirical proportion of affected objects after observing *Some of the X Yed*, as a function of prior mean proportion of affected objects.

sults of three experiments (1, 2a, 2b) that show that while there is an effect of the prior on listeners’ interpretations of sentences like *Some of the marbles sank*, this effect is much smaller than predicted by rRSA. Exp. 3 provides evidence that listeners’ beliefs about object or event wonkiness are indeed influenced by the surprisal of the utterance. Finally, we present wRSA as an extension of rRSA that incorporates the idea of backing off to alternate prior beliefs if the observed utterance suggests a wonky world. This model provides a much better fit to the empirical data from Exps. 2a and 2b than rRSA, and also provides a good fit to the wonkiness ratings obtained in Exp. 3.

Include also plot of all-state probability: posterior empirical and predicted probability (rRSA, wRSA) as a function of prior all-state probability, for the SI people

## Experiment 1

Exp. 1 measured  $P(s)$ , participants’ prior beliefs about different objects exhibiting a certain effect (e.g., marbles sinking).

## Method

**Participants** We recruited 60 participants over Amazon’s crowd-sourcing platform Mechanical Turk.

**Procedure and materials** On each trial,<sup>2</sup> participants read a description of an event like *John threw 15 marbles into a pool*. They were then asked to provide a judgment of an effect, e.g. *How many of the marbles do you think sank?*, on a sliding scale from 0 to 15. Judgments were obtained for 90 items, of which each participant saw a random selection of 30 items.

<sup>2</sup>This experiment can be viewed at [https://www.stanford.edu/~jdegen/12\\_sinking-marbles-prior15/sinking-marbles-prior.html](https://www.stanford.edu/~jdegen/12_sinking-marbles-prior15/sinking-marbles-prior.html)

## Results

Data from one participant, who gave only one response throughout the experiment, were excluded. Each item received between 12 and 29 ratings. Distributions of ratings for each item were smoothed using nonparametric density estimation for ordinal categorical variables (Nelson, 1989) using the np package in R (Nelson, 1989) see hayfield 2013 np package specification for li and racine ref.

### Experiment 2a

Exp. 2a measured  $p(s|u)$ , participants' posterior beliefs in different objects exhibiting a certain effect (e.g., marbles sinking), after observing an utterance.

**Participants** We recruited 120 participants over Amazon's crowd-sourcing platform Mechanical Turk.

#### Procedure and materials <sup>3</sup>

Participants read the same descriptions as in Exp. 1. They additionally saw an utterance produced by a knowledgeable speaker about the event, e.g. *John, who observed what happened, said: "Some of the marbles sank"*, and were asked to rate on sliding scales with endpoints labeled "very unlikely" and "very likely", how likely they thought 0%, 1-50%, 51-99%, or 100% of the marbles sank.

Each participant saw 10 "some" trials and 20 fillers, of which 10 contained the quantifiers "all" or "none", and the rest were utterances that did not address the number of objects that displayed the effect, e.g. *What a stupid thing to do*.

#### Results XXX question

$p(s_v|u_{\text{some}})$  increased with increasing talk about both  $p(s_v)$  and the prior expectation of the distribution?  $p(s_v)$  ( $\beta=.1$ ,  $SE=.01$ ,  $t=6.9$ ,  $p<.0001$ ); however, mean  $p(s_v|u_{\text{some}})$  was never higher than .26, suggesting that a) participants drew strong implicatures in this paradigm and b) the effect of  $P(s)$  is much smaller than predicted by rRSA.

### Experiment 2b

Exp. 2b replicates the effect of the prior on participants' posterior estimates of different objects exhibiting a certain effect (e.g., marbles sinking) using a different dependent measure.

**Participants** We recruited 120 participants over Amazon's crowd-sourcing platform Mechanical Turk.

#### Procedure and materials <sup>4</sup>

The procedure and materials were identical to those of Exp. 2a with the exception of the dependent measure. Rather than providing point estimates of the probability of different numbers of objects sinking, participants performed the task from Exp. 1, i.e., they were asked to provide a judgment of

<sup>3</sup>This experiment can be viewed at <https://web.stanford.edu/~jdegen/sinking-marbles-nullutterance/sinking-marbles-nullutterance.html>

<sup>4</sup>This experiment can be viewed at [https://www.stanford.edu/~jdegen/13\\_sinking-marbles-prior-dv-15/sinking-marbles.html](https://www.stanford.edu/~jdegen/13_sinking-marbles-prior-dv-15/sinking-marbles.html)

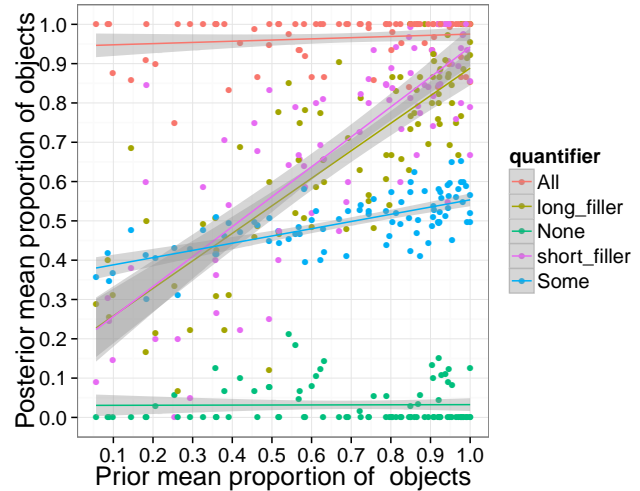


Figure 2: For each item, mean empirical proportion of affected objects after observing an utterance, as a function of prior mean proportion, for different quantifiers and filler conditions.

an effect, e.g. *How many of the marbles do you think sank?*, on a sliding scale from 0 to 15.

#### Results and discussion XXX question

The mean number of objects judged to exhibit the effect increased with increasing expectation of the prior distribution ( $\beta=.18$ ,  $SE=.02$ ,  $t=7.4$ ,  $p<.0001$ , see also Figure ??), replicating the effect observed in Exp. 2a. Again, the effect of the prior was much smaller than predicted by rRSA and resulted in mean proportions of affected objects between 30% and 65%, where rRSA predicts a range from XXX to XXX for these items.

Exps. 2a and 2b demonstrate that there is a robust effect of listeners' prior beliefs on the interpretation of utterances with *some*. However, this effect is much smaller than predicted by rRSA.

#### Discussion of why; wonky world intuition

### Experiment 3

Exp. 3 measured participants' beliefs in world wonkiness after observing the scenarios and utterances from Exps. 2a and 2b.

**Participants** We recruited 60 participants over Amazon's crowd-sourcing platform Mechanical Turk.

#### Procedure and materials <sup>5</sup>

The procedure and materials were identical to those of Exps. 2a and 2b, with the exception of the dependent measure. Rather than providing estimates of what they believed the world was like, participants were asked to indicate how

<sup>5</sup>This experiment can be viewed at [https://web.stanford.edu/~jdegen/17\\_sinking-marbles-normal-sliders/sinking-marbles-normal.html](https://web.stanford.edu/~jdegen/17_sinking-marbles-normal-sliders/sinking-marbles-normal.html)

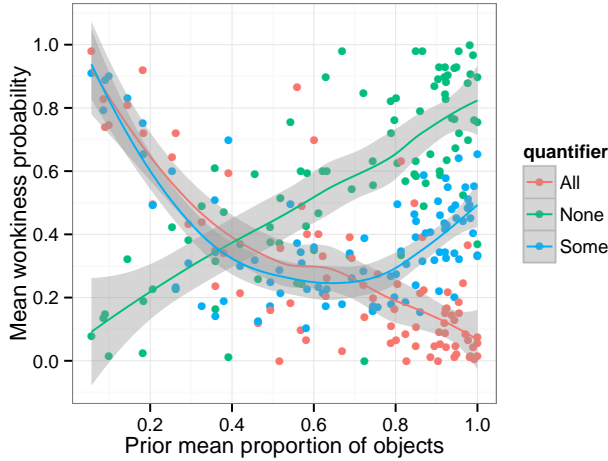


Figure 3: For each item, mean wonkiness probability after observing an utterance, as a function of expected prior proportion of affected objects, for different quantifiers.

likely it was that the objects (e.g., the marbles) involved in the scenario were normal objects, by adjusting a slider that ranged from *definitely not normal* to *definitely normal*.

**Results** The extreme ends of the sliders were coded as 1 (*definitely not normal*, i.e., wonky) and 0 (*definitely normal*, i.e., not wonky). We interpret the slider values as probability of world wonkiness. Mean wonkiness probability ratings are shown in Figure ?? . For *all* and *none*, increasing prior expectation of objects exhibiting the effect resulted in a fairly linear decrease and increase in the probability of wonkiness, respectively. For *some*, the pattern is somewhat more intricate: probability of wonkiness initially decreases sharply, but rises again in the upper range of the prior expected value.

yay!! XXX

### Model

The wonky RSA model we propose for capturing the defeasibility of listeners' world knowledge is an extension of regular RSA: in wRSA, the listener infers the value of  $\theta_{\text{wonky}}$  jointly with  $s$ .  $\theta_{\text{wonky}}$  captures for each utterance and item, how likely the objects involved in the event (e.g., marbles) are in fact "wonky" (in which case the computation draws on a uniform prior, i.e. disregards prior beliefs) or not (in which case the model draws on the smoothed empirical prior distribution for that item, obtained in Exp. 1). The resulting  $p(s|u)$  is a mixture of computations based on the uniform and empirical prior, with mixture parameter  $\theta_{\text{wonky}}$ . The inferred value of  $\theta_{\text{wonky}}$  itself depends on  $p(u|s)$ : the more surprising a particular utterance is given prior beliefs, the higher the probability of  $\theta_{\text{wonky}}$ .

### Model evaluation

From the abstract: Comparing the fit of rRSA and wRSA model predictions to the posterior state estimates from Exp. 2

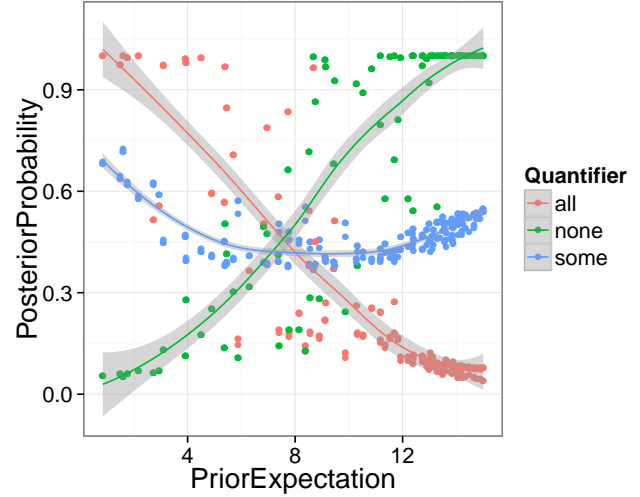


Figure 4: For each item, predicted proportion of 'wonky' ratings after observing an utterance, as a function of prior mean proportion, for different quantifiers.

yields a much better fit for wRSA. The better fit of wRSA suggests that listeners use speakers' utterances as cues to how strongly to incorporate world knowledge. wRSA also provided a better fit than a model which used only a uniform prior, confirming that listeners do make use of world knowledge in a systematic way in the computation of scalar implicature.

### General discussion

- what is wonky?
- other ways of asking about wonkiness
- what's the right prior to back off to?

### Conclusion

### References

- Frank, M. C., Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336, 998.
- Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge Univ Press.