Wonky worlds: Listeners revise world knowledge when utterances are odd

Judith Degen, Michael Henry Tessler, Noah D. Goodman

{jdegen,mhtessler,ngoodman}@stanford.edu

Department of Psychology, 450 Serra Mall

Stanford, CA 94305 USA

April 10, 2015

Abstract

World knowledge enters into pragmatic utterance interpretation in complex ways, and may be defeasible in light of speakers' utterances. Yet there is to date a surprising lack of systematic investigation into the role of world knowledge in pragmatic inference. In this paper, we show that a state-of-the-art model of pragmatic interpretation greatly overestimates the influence of world knowledge on the interpretation of utterances like *Some of the marbles sank*. We extend the model to capture the idea that the listener is uncertain about the background knowledge the speaker is bringing to the conversation. This extension greatly improves model predictions of listeners' interpretation and also makes good qualitative predictions about listeners' judgments of how 'normal' the world is in light of a speaker's statement. Theoretical and methodological implications are discussed.

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

1 Introduction

How often do you think marbles sink in water? Probably extremely often, if not always. Now imagine someone says *Max threw fifteen marbles in the water. Some of the marbles sank.* Have you

begun to reconsider your assumptions? Perhaps you now suspect that these marbles are in fact made of hollow plastic or the water is covered with thick algae? That is, maybe you have begun to suspect that these are not just normal marbles in normal water. Alternatively, maybe you know the speaker who produced the utterance and know them to be unreliable in their descriptions of the world. Here we explore how prior world knowledge enters into pragmatic utterance interpretation, and when this world knowledge is defeasible: some utterances are odd and may lead listeners to conclude that the world under discussion is abnormal in the sense that events in that world have appropriately different prior probabilities. We refer to such abnormal worlds as *wonky* worlds. However, sometimes the oddness of those same utterances may be explained away by the knowledge of *speaker unreliability*, in which case listeners' prior beliefs about the world that they bring to bear on the utterance situation should remain unchanged.

The Rational Speech Act framework (RSA, Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013), and related models (Franke, 2011; Russell, 2012), treat communication as a signaling game (Lewis, 1969) between a speaker and a listener. The listener reasons by Bayesian inference about what the world is like given that a speaker who produced the utterance is trying to be informative (with respect to a naïve listener who interprets utterances literally). Variants of these models have successfully captured listeners' quantitative behavior on a number of pragmatic inference tasks, including ad hoc Quantity implicature (Degen, Franke, & Jäger, 2013), markedness implicature (Bergen, Goodman, & Levy, 2012), scalar implicature (Goodman & Stuhlmüller, 2013), syllogistic reasoning (?, ?), and non-literal language (Kao, Wu, Bergen, & Goodman, 2014). A defining feature of Bayesian reasoning is that prior beliefs affect inferences that will be drawn. Bayesian models of language interpretation, accordingly, predict that prior beliefs about the world should affect the listener's interpretation of an utterance. While this impact of prior knowledge has been noted, and included in models, it hasn't been systematically studied.

Generalizing our opening example, consider *Some of the X sank*, where X is a plural noun such as *marbles*, *feathers*, or *balloons*, and *the X* refers to a contextually established group of objects from category X. When the prior probability, θ_X, of an X[1] sinking is not extreme (e.g., a feather

[1]We will use 'X' interchangeably to refer to both the category and the members of the category.

sinking), RSA leads to the standard scalar implicature: the posterior probability that all of the X sank, after hearing the utterance, is much lower than its prior probability (i.e., *Some of the feathers sank* yields that not all of them did). This is because a rational speaker would have been expected to produce the more informative (and contextually relevant) *All of the X sank*, had it been true. As we will show in Section 3, RSA makes two strong predictions about the effect of the prior on the interpretation of *Some of the X sank*:

1. As θ_X approaches 1, the interpretation probability that all X sank approaches 1, that is, the scalar implicature disappears. This prediction follows because the extreme prior overwhelms the effect of the utterance's semantics.

2. For moderate to high prior probability (roughly $0.5<\theta_X<1$) and a large total number of objects (more than about 10), the posterior expectation of the number of X that sank should be approximately the same as the prior expectation—that is, the utterance shouldn't affect the expected number of X that sank.

These predictions follow from the weak semantics of *some* and the isolated effect of the alternative *all*: *Some of the X sank* only restricts the interpretation (i.e., the number of X that sank) to be greater than zero; competition with *All of the X sank* results in the scalar implicature that can at most rule out the state in which all of the X sank. But that leaves at least fourteen other possibilities: that one X sank, that two X sank, etc. Thus, a sufficiently strong prior will dominate the inference about exactly how many X sank.

However, intuition is at odds with these predictions: as Geurts (2010) has observed, for events with very high prior probability of occurrence (e.g., marbles sinking), observing an utterance such as *Some of the marbles sank* seems to yield strong implicatures; that is, contrary to RSA predictions, the subjective probability that all of the marbles sank is intuitively close to 0.

The rest of the paper is structured as follows. In Exp. 1 we collect prior probabilities for a variety of events (e.g., sinking) and categories that participate in those events (e.g., marbles). In Exps. 2a and 2b we collect corresponding posterior interpretations after observing descriptions of those events containing quantifiers (*some*, *all*, *none*). These experimental results confirm the

intuition of relatively strong implicatures—hence prediction (1) of RSA is incorrect—and show that the prior has a muted effect on posterior expectation—hence prediction (2) of RSA is incorrect.

Given the previous success of RSA models, this constitutes a striking puzzle. To address this puzzle 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 use less extreme prior beliefs in the computation of speaker meaning. In Section 5 we introduce a variant of RSA, *wonky RSA (wRSA)*, in which the listener can revise her beliefs about the domain under discussion. We show that this extension resolves the puzzle of the prior's muted effects.

The extended wRSA model makes predictions not just about the actual state of the world that the listener infers the speaker as trying to communicate, but also about listeners' judgments of world wonkiness. This allows for a further test of wRSA: in Exp. 3 we collect participants' judgments about whether the world is wonky in the scenarios of Exps. 2a and 2b.

Finally, in Section 7 we test the prediction that listeners should hold on to their prior beliefs when interpreting odd utterances if they expect speakers to be unreliabile. That is, if utterance oddness can be explained away by speaker unreliability, prior beliefs about how many X Y should have a greater effect on the interpretation of *Some of the X Yed* than when speakers are expected to be reliable. We test the effect of speaker reliability on interpretation of a subset of the items used in Exps. 2a-3 in Exp. 4. In Section 8 we discuss the results with respect to a) ideal adaptor models in other areas of psycholinguistics and b) presupposition accommodation / common ground update.

2 Experiment 1a and 1b: prior elicitation

Obtaining good estimates of prior beliefs is crucial for testing any Bayesian model. Indeed, Bayesian approaches have recently been criticized for being lax in their treatment of priors (Jones & Love, 2011; Marcus & Davis, 2013). To avoid this criticism, we would thus like to obtain good empirical estimates of prior beliefs. A problem in this endeavor is that there are no clearly established dependent measures for eliciting prior beliefs about different types of events. One way to solve

this problem is to elicit priors using different dependent measures and using the data obtained in this way to infer the underlying 'true' prior. This is what we do here. Exp. 1a and 1b[2] measured listeners' prior beliefs about how many objects exhibit a certain effect (e.g., how many marbles sink) using two different dependent measures. The 'true' prior that is used in all subsequent analyses throughout the paper was then inferred by XXX MH summarize procedure in one sentence.

2.1 Method

2.1.1 Participants

For each experiment, we recruited 60 participants over Amazon's crowd-sourcing platform Mechanical Turk. Participants were paid \$0.50 (Exp. 1a) and 2\$ (Exp. 1b), respectively, for their participation. Here and in all other experiments reported in this paper, participants' IP address was limited to US addresses only and only participants with a past work approval rate of at least 95% were accepted.

2.1.2 Procedure and materials

On each trial, participants read a one-sentence 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?*. In Exp. 1a, they chose a number between 0 and 15 to indicate how many marbles they thought sank. In Exp. 1b, they instead rated for each number of marbles from 0 to 15, how likely they thought that number of marbles sank, by adjusting a slider with endpoints labeled "impossible" and "certain".

Each item had a similar form: the first sentence introduced the objects at issue (e.g., marbles). The question always had the form *How many of the X Yed?*, where *X* was the head of the direct object noun phrase introduced in the first sentence (e.g., *marbles*, *cups*, *balloons*) and *Yed* was a verb phrase denoting an effect that the objects underwent (e.g., *sank*, *broke*, *stuck to the wall*). Each verb phrase occurred with three different objects, e.g., *sank* occurred with *marbles*, *cups*,

[2]These experiments can be viewed at `http://cocolab.stanford.edu/cogsci2015/wonky/prior/sinking-marbles-prior.html` and `http://cocolab.stanford.edu/cogsci2015/wonky/prior_sliders/sinking-marbles.html`.

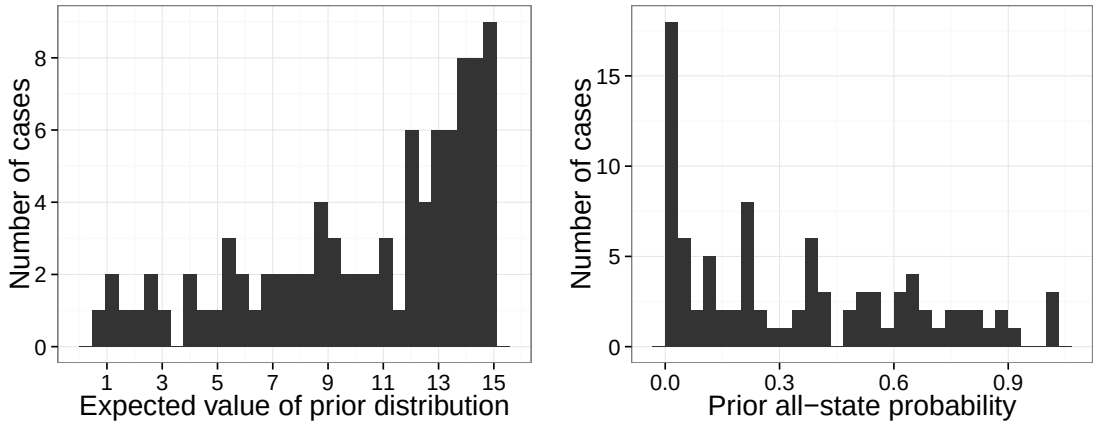

Figure 1: Histogram of expected values $\mathbb{E}[P(s)]$ of each empirically elicited and smoothed prior distribution (left) and histogram of prior probabilities $P(s_{15})$ of the all-state for each item (right).

and *balloons*. Items were constructed to intuitively cover the range of probabilities as much as possible, while also somewhat oversampling the upper range of probabilities to have more fine-grained coverage of this region that is of most interest for testing the RSA model. Judgments were obtained for 90 items, of which each participant saw a random selection of 30 items.

2.2 Results

Data from one participant in Exp. 1a, who gave only one response throughout the experiment, were excluded. Each item received between 12 and 29 ratings in each sub-experiment. XXX MH insert procedure on prior inference; discuss the results from the two different dependent measures and how we expect one to over-sample extremes and one to be too flat? As intended, items covered a wide range of probabilities. See Figure 1 for a histogram of expected values of each smoothed prior distribution as well as prior all-state probabilities for each item.

In the next section, we use the underlying prior beliefs inferred from the two different dependent measures to derive RSA predictions for the interpretation of utterances like *Some of the marbles sank*, before empirically measuring participants' interpretations.

3 Effect of the world prior in RSA

The basic Rational Speech Acts model defines a pragmatic listener $P_{L_1}(s|u)$ who reasons about a speaker $P_{S_1}(u|s)$, who in turn reasons about a literal listener $P_{L_0}(s|u)$. Each listener performs Bayesian inference about the world state the speaker intends to communicate, given either the literal truth of utterance u or the speaker's choice of u; the speaker is a softmax-optimal decision maker, with the goal of being informative about the state s. RSA is defined by:

$$P_{L_0}(s|u) \propto \delta_{[\![u]\!](s)} \cdot P(s) \tag{1}$$

$$P_{S_1}(u|s) \propto \exp(\lambda \ln P_{L_0}(s|u)) \tag{2}$$

$$P_{L_1}(s|u) \propto P_{S_1}(u|s) \cdot P(s) \tag{3}$$

Here $[\![u]\!] : S \rightarrow$ Boolean is a truth-function specifying the literal meaning of each utterance and $\delta_{[\![u]\!](s)}$ is the Kronecker delta function returning a uniform distribution over all states s compatible with utterance u.

For concreteness, assume that the set of states of the world is $S = \{s_0, s_1, s_2, \ldots, s_{15}\}$, where the subscript indicates the number of objects (e.g., marbles) that exhibit an effect (e.g., sinking). Further assume that the set of three utterances *All/None/Some of the marbles sank* is denoted $U = \{u_{\text{all}}, u_{\text{none}}, u_{\text{some}}\}$ and each has its usual literal meaning: $[\![u_{\text{none}}]\!] = \{s_i | i = 0\}$, $[\![u_{\text{some}}]\!] = \{s_i | i > 0\}$, $[\![u_{\text{all}}]\!] = \{s_i | i = 15\}$.

In Figure 2 we show the predictions of RSA (dark blue dots) for the items from Exps. 1a and 1b in two different ways: the left panel shows the posterior expected number of affected objects as a function of the prior expectation; the right panel shows the posterior probability of the state in which all objects are affected, as a function of the prior probability of that state.[3] We see that the prior has a strong effect, which can be summarized by the two predictions described in the Introduction:

1. $P(s_{15}|u_{\text{some}}) \rightarrow 1$ as $P(s_{15}) \rightarrow 1$

[3] That the individual model predictions look somewhat noisy is due to the different shapes of the prior distributions, such that for the same expected value of the distribution, the distribution itself can take different shapes, which are treated slightly differently by the model.

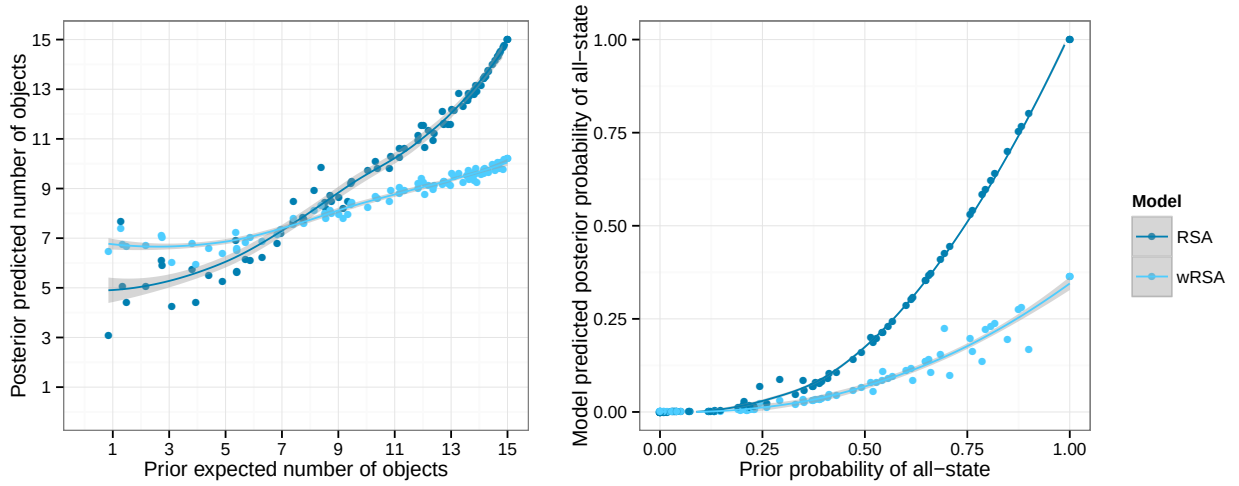

Figure 2: For each item, RSA and wRSA model predicted $\mathbb{E}[P(s|u_{\text{some}})]$ as a function of $\mathbb{E}[P(s)]$ (left) and $P(s_{15}|u_{\text{some}})$ as a function of $P(s_{15})$ (right).

2. $\mathbb{E}[P(s|u_{\text{some}})] \simeq \mathbb{E}[P(s)]$ over the upper half of its range.

We next turn to an empirical test of these predictions, or rather, of the intuition that they may be incorrect.

4 Experiment 2a and 2b: comprehension

Exps. 2a and 2b[4] measured participants' posterior beliefs $P(s|u)$ about how many objects exhibited a certain effect (e.g., marbles sinking), after observing an utterance. The only difference between the experiments was the dependent measure. The dependent measures differed in order to directly and independently estimate the two values that the predictions above are concerned with: $\mathbb{E}[P(s|u_{\text{some}})]$ and $P(s_{15}|u_{\text{some}})$, i.e., the expected number of X that are deemed to have Yed and the probability of all of the X having Yed (a measure of implicature strength), respectively, after observing the utterance *Some of the X Yed*.

[4]These experiments can be viewed at `http://cocolab.stanford.edu/cogsci2015/wonky/expectation/sinking-marbles.html` and `http://cocolab.stanford.edu/cogsci2015/wonky/stateprobs/sinking-marbles-nullutterance.html`

4.1 Method

4.1.1 Participants

For each experiment we recruited 120 participants over Amazon's Mechanical Turk who were paid \$0.70 for their participation.

4.1.2 Procedure and materials

Participants read the same descriptions as in Exps. 1a and 1b. 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"*. In Exp. 2a (just as in Exp. 1a), they then provided a judgment of an effect, e.g. *How many of the marbles do you think sank?*, on a sliding scale from 0 to 15. In Exp. 2b they instead rated on sliding scales with endpoints labeled "definitely not" and "definitely", how likely they thought 0%, 1-50%, 51-99%, or 100% of the objects exhibited the effect.

Each participant saw 10 *some* trials and 20 filler trials, 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. These 10 additional fillers were intended to establish a baseline for participants' use of information about the prior. Of these, half were generic short fillers that were intended to communicate the prior, e.g., *Typical*. The rest were longer sentences that addressed a different aspect of the described scenario, e.g. *What a stupid thing to do*.[5] The utterances were randomly paired with 30 random items for each participant.

4.2 Results and discussion

Data from Exp. 2a were analyzed as given. Slider ratings obtained in Exp. 2b were first subjected to normalization on a by-trial and by-participant basis, such that on each trial, a participant's ratings added to 1, i.e., constituted a proper probability distribution. After normalization, data from eight participants in Exp. 2b were excluded from the analysis because these participants assigned less than 0.8 probability to the interpretation corresponding to the correct literal interpretation on

[5]See Appendix A for a complete list of items.

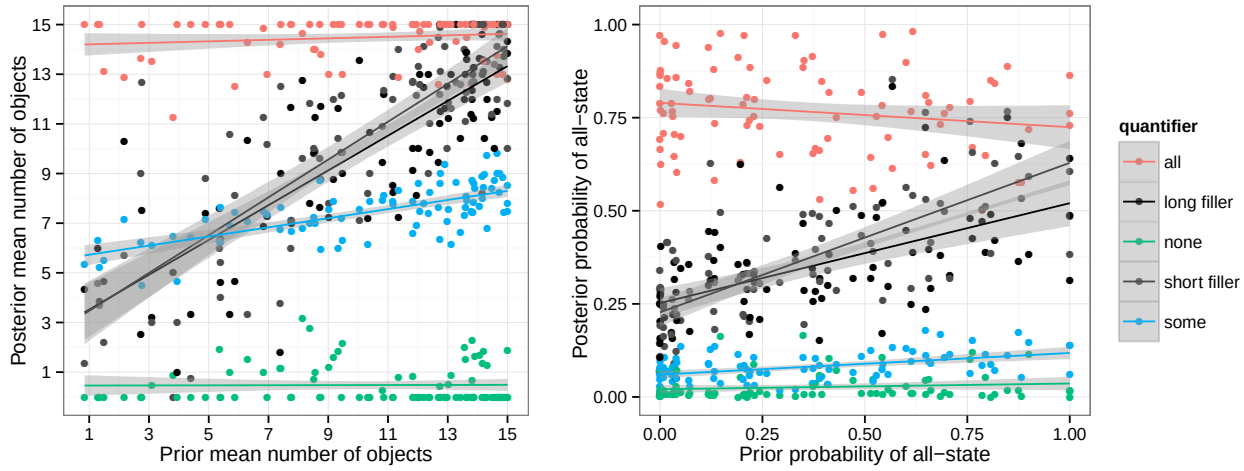

Figure 3: For each item and quantifier, empirical $\mathbb{E}[P(s|u_{\text{some}})]$ from Exp. 2a against $\mathbb{E}[P(s)]$ from Exp. 1 (left) and empirical $P(s_{15}|u_{\text{some}})$ from Exp. 2b against $P(s_{15})$ from Exp. 1 (right).

literal *all* and *none* trials.[6]

The main question of interest was whether the predictions of the basic RSA model laid out in the previous section were borne out in participants' judgments of how many X Yed after hearing an utterance with *some*. Mean empirical $\mathbb{E}[P(s|u)]$ (expected value of the posterior distribution) and $P(s_{15}|u)$ (posterior all-state probability) are shown in Figure 3 for each item. Visual inspection of the graphs shows that the interpretation of utterances with *all* and *none* seems to be relatively unaffected by the prior, while the interpretation of *some* displays a small, but robust, effect.

To test whether the visual effect of the prior on the interpretation of *some* is real, we conducted linear mixed effects regressions on the *some* data from each experiment. For Exp. 2a, the number of X that Yed was regressed onto centered fixed effects of each item's prior expectation and trial number (to account for changes in response behavior over the course of the experiment). For Exp. 2b, the all-state probability was regressed onto centered fixed effects of each item's prior all-state probability and trial number. Each model also contained the maximal by-item and by-participant random effects structure, following the guidelines outlined by Barr, Levy, Scheepers, and Tily (2013). Rerun all the analyses once you get numbers from MH.

[6]In general, this task yielded noisier results than the task in Exp. 2a (as can be seen in the average lower probability of the all- state after observing *all*, in the right panel of Figure 3) because participants used the sliders in different ways. For example, for cases where intuitively, the all-state was true, some participants assigned non-zero probability to only the all-state, while others were reluctant to do so and always assigned some probability to the 51-99% state.

For utterances of *Some of the X Yed*, the mean number of objects judged to exhibit the effect increased with increasing expectation of the prior distribution (β=.18, SE=.02, t=7.4, p<.0001). Similarly, the probability of all 15 objects exhibiting the effect increased with increasing prior probability of doing so (β=.06, SE=.01, t=5.0, p<.0001). However, the size of these effects is, to say the least, much smaller than predicted by RSA (for comparison, see dark lines in Figure 2).

One possible explanation for this highly attenuated effect of the prior is that participants simply did not bring world knowledge to bear on the interpretation of utterances. However, this possibility is ruled out by examining participants' performance in the filler conditions: in both Exps. 2a and 2b, the filler conditions closely tracked the prior (see Figure 3). Additionally, that there is any effect of the prior at all suggests that participants are not entirely disregarding their prior beliefs.

Exps. 2a and 2b thus demonstrate that there is an effect of listeners' prior beliefs on the interpretation of utterances with *some*. However, this effect is quantitatively much smaller than predicted by RSA, and qualitatively does not match the predictions identified above: the implicature is not canceled for extreme priors (contra prediction (1)) and the posterior expectation diverges from the prior expectation (contra prediction (2)). In the next section, we extend the RSA model to formalize the intuition, raised in the Introduction, that a listener may decide that her initial beliefs about the domain are not shared by the speaker and respond by revising her priors.

5 Effect of the world prior in 'wonky RSA'

To capture the idea that the pragmatic listener is unsure what background knowledge the speaker is bringing to the conversation, we extend the basic RSA model by using a "lifted variable" (Goodman & Lassiter, in press; Lassiter & Goodman, 2013; Bergen et al., 2012; Kao et al., 2014) corresponding to the choice of state prior. A lifted variable is a variable that the model (in particular, the pragmatic listener) reasons about explicitly instead of being given a value for it. In this case, we posit that the prior, now $P(s|w)$, depends on a "wonkiness" variable w, which determines whether to use the

"usual" prior for this domain or a more generic back-off prior, which we take to be uniform:[7]

$$P(s|w) \propto \begin{cases} 1 & \text{if } w \\ P_{\text{usual}}(s) & \text{if not } w \end{cases}$$

This inferred prior is used in both the literal and pragmatic listeners, indicating that it is taken to be common ground. However, the w variable is reasoned about only by the pragmatic listener, which captures the idea that it is an inference the pragmatic listener makes about which assumptions are appropriate to the conversation. Using the notation of Section 3:

$$P_{L_0}(s|u,w) \propto \delta_{[\![u]\!](s)} \cdot P(s|w) \tag{4}$$

$$P_{S_1}(u|s,w) \propto \exp(\lambda \ln P_{L_0}(s|u,w)) \tag{5}$$

$$P_{L_1}(s,w|u) \propto P_{S_1}(u|s,w) \cdot P(s|w) \cdot P(w) \tag{6}$$

We refer to this model as wRSA. Notice that the choice of w that the listener makes will depend on $P_{S_1}(u|s,w)$: if a given utterance can't be explained by the usual prior, because it is unlikely under any plausible world state s, then the pragmatic listener will infer that the world is wonky, and back off to the uniform prior. That is, if the utterance is odd, the listener will revise her opinion about what world knowledge is appropriate to use. we need to explain where oddness comes from – maybe there's an easy graphical way of showing how for items with different priors, a 'some' utterance is more or less wonky, by showing the marginal probabilities of observing each utterance for these different items?

update this bit with mh's numbers To make predictions for Exps. 2 from wRSA we use the inferred priors from Exps. 1 as $P_{\text{usual}}(s)$ for each item. The wonkiness prior $P(w)$ and the speaker optimality parameter λ are fit to optimize mean squared error (MSE) with Exp. 2 data. The optimal parameters ($\lambda = 2$, $P(w) = 0.5$) resulted in an MSE of 2.15 (compared to 14.53 for RSA) for the expected number of objects, and 0.01 (compared to 0.07 for RSA) for the all-state probability. The better fit of wRSA compared to RSA can be seen in the comparison of Figure 2 and Figure 3: in

[7] a footnote either briefly discussing other priors or delaying such a discussion to the general discussion

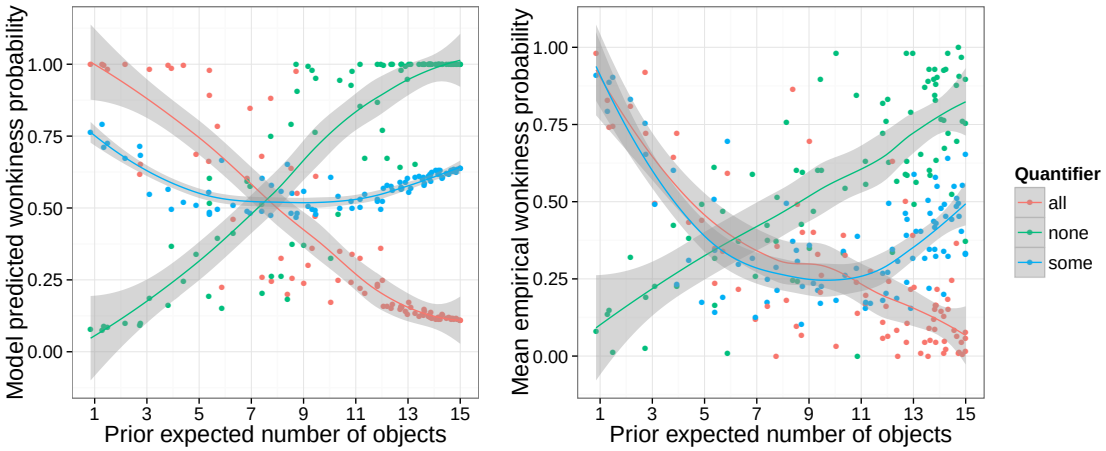

Figure 4: For each item, predicted (left) and empirical (right) wonkiness probability after observing an utterance ($u_{\text{all}}, u_{\text{none}}, u_{\text{some}}$), as a function of the prior expected number of affected objects.

both cases, wRSA (light blue lines) predicts a much attenuated effect of the prior compared to regular RSA (dark blue lines), in line with the empirical data. Furthermore, wRSA does not make either of the problematic predictions identified earlier for regular RSA.

These results are encouraging: wRSA is able to account for the qualitative and quantitative departures of participants' behavior from RSA, with respect to the effect of the prior. Is this because listeners are actually inferring that the world is unusual from an utterance like *Some of the marbles sank*? The wRSA model makes predictions about the probability that a given world is wonky after observing an utterance; see the left panel of Figure 4 for predicted wonkiness probabilities for u_{all}, u_{none}, and u_{some} using the optimal $P(w)$ and λ parameters from fitting wRSA to the Exp. 2 data. Note the U-shaped curve, in which the world is judged wonky if u_{some} is used in worlds with extreme priors. We can test these predictions directly by simply asking participants for each item from Exps. 2a and 2b, whether they believe the described situation is normal.

6 Experiment 3: wonkiness

Exp. 3[8] measured participants' beliefs in world wonkiness after observing the scenarios and utterances from Exps. 2a and 2b.

[8]This experiment can be viewed at `http://cocolab.stanford.edu/cogsci2015/wonky/wonkiness/sinking-marbles-normal.html`

6.1 Methods

6.1.1 Participants

We recruited 60 participants over Amazon's crowd-sourcing platform Mechanical Turk who were paid \$0.50 for their participation.

6.1.2 Procedure and materials

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 how many X they believed Yed, participants were asked to indicate how likely it was that the objects (e.g., the marbles) involved in the scenario were normal objects, by adjusting a slider with endpoints labeled "definitely not normal" to "definitely normal."

6.2 Results and discussion

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 the right panel of Figure 4 and closely mimic wRSA's predictions (see left panel of Figure 4). For u_{all} and u_{none}, increasing prior expectation of Xs Ying resulted in a fairly linear decrease and increase in the probability of wonkiness, respectively. For u_{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.

Qualitatively, the model captures both the linear increase and decrease in wonkiness probability for u_{all} and u_{none}, respectively. Importantly, it also captures the asymmetric U-shaped wonkiness probability curve displayed by u_{some}. Intuitively, this shape can be explained as follows: for very low probability events, it is surprising to learn that such an event took place (which is what is communicated by u_{some}), so wonkiness is high. For medium probability events, learning that this event took place is not very surprising, so wonkiness is relatively low. For high probability events, u_{some} may be literally true, but it is not useful in the sense of providing the listener new information.

For comparison to the comprehension data fit, the model's MSE for empirical wonkiness probability predictions, using the best parameters from fitting the model to the comprehension data, was 0.07.

That the wonkiness probability predictions are borne out in the empirical data provide further support for wRSA, and for the idea that participants are revising their prior beliefs online when encountering an odd utterance. put in here somewhere the relation between wonkiness and the comprehension data: for all and none, while there are huge changes in wonkiness by prior, we don't expect this to show up in the comprehension data because the semantics of the utterances restricts the interpretation to just one state, regardless of the prior. but for "some", which has a weak semantics, wonkiness shifts the overall interpretation in a way that compresses the effect of the prior

Some readers may still not be convinced that what is driving the attenuated effect of the prior is that listeners revise their beliefs about the world (or about the event, or the objects involved in the event—we return to this issue in the general discussion). A prima facie alternative possibility to why prior beliefs have a much smaller effect than expected on utterance interpretation is that listeners believe that the speaker is an unreliable or uncooperative speaker, one who talks about the world in surprising and unexpected ways. For example, maybe participants in the studies reported thus far believed that the speaker was a) lying, b) not trying to be informative, c) or not actually knowledgeable about the precise state of the world, i.e., about how many X Yed.

There are multiple reasons why this is not a plausible explanation for our results. First, participants were happy to interpret the speaker's utterances literally in the *all* and *none* conditions, suggesting they didn't believe the speaker was lying, ruling out a). Second, the wonkiness judgments obtained in Exp. 3 closely match the wRSA model's predictions, which relies on speaker reliability and cooperativity—knowing the actual state of the world and reporting it truthfully and informatively—, ruling out b) and c). However, there is an even clearer reason for ruling out speaker reliability explanations for the observed attenuated effect of the prior: if the speaker was actually unreliable, intuition and the (RSA and wRSA) model predict that listeners should use their prior world knowledge much more strongly than if the speaker was reliable. For example, if my very successful but understated friend says that he did "OK" on a job interview, I'll be more

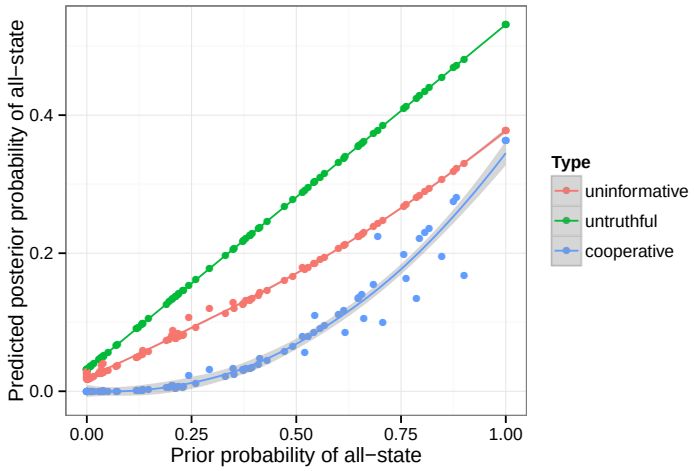

Figure 5: For each item, predicted mean posterior all-state probability against prior all-state probability, for uninformative, untruthful, and cooperative speakers. (Note that the cooperative speaker predictions are repeated from Figure 2).

likely to disregard his pronouncement of mediocre performance and expect him to have done very well. That is, speaker unreliability (however well intentioned) intuitively leads interlocutors to much more strongly rely on their prior beliefs and discount the speaker's utterance.

From the model's perspective, we can think about speaker unreliability in terms of different ways of lesioning the model. Recall that the speaker chooses utterances in proportion to the informativeness of that utterance to a literal listener. One way to lesion the model is thus to let the speaker produce a random true utterance (as opposed to an informative true utterance). The other way is to let the speaker produce an utterance at random from the set of alternatives, regardless of whether it is true. This results in differences in the pragmatic listener's interpretation probabilities. In particular, the truthful but uninformative speaker somewhat elevates the effect of the prior on the posterior all-state probability in wRSA, and the neither truthful nor informative speaker further increases that effect, just as intuition suggests. These model predictions are shown for the items used in Exps. 1-3 in Figure 5, using the best parameters ($\lambda = 2$, $P(w) = .5$) from the model fitting reported in Section 5.

These considerations make the following prediction: if speaker reliability is explicitly manipulated, listeners' interpretations of *Some of the X Yed* should be much more strongly governed by

their prior beliefs when the speaker is unreliable (either untruthful or uninformative) than when the speaker is reliable/cooperative. That is, we expect a replication of Exp. 2b for reliable, but not unreliable speakers. If, instead, there is something more fundamentally wrong with RSA models and the attenuated effect of the prior on comprehension is driven by listeners' perception of the speaker as unreliable, further decreasing speaker reliability should have no effect on how strongly the prior influences listeners' interpretations. If anything, explicitly unreliable speakers should further diminish the effect of the prior. We test the prediction about speaker reliability in Exp. 4.

7 Exp. 4: speaker reliability

Exp. 4[9] explores how speaker reliability interacts with prior beliefs in the interpretation of the same utterances as in Exps. 1-3. As discussed in the previous section, the wRSA model predicts that listeners' interpretation of *Some of the X Yed* should be more strongly affected by their prior beliefs when the speaker is perceived as unreliable than when the speaker is perceived as reliable. There are different ways in which speaker reliability can be construed. Here, we are interested in manipulating speaker reliability in such a way as to lead listeners to expect underinformativity or truthful (but wrong) reports of misperceived events. To this end, we introduce two unreliable speakers: a speaker who has an incentive to be misleading in a courtroom scenario and a drunk speaker. These speakers are contrasted within participants with a reliable speaker.

Exp. 4 was very similar to Exp. 2b, but used only half of the items and a blocked design. In each block, the speaker who described observed events remained the same throughout the block. The speaker's reliability was established via a short cover story at the beginning of each block. This allowed us to compare the effect of prior beliefs on participants' interpretations of *Some of the X Yed* uttered by speakers of differing reliability.

[9]This experiment can be viewed at `http://cocolab.stanford.edu/cogsci2015/wonky/speakerreliability/sinking-marbles.html`.

7.1 Method

7.1.1 Participants

We recruited 120 participants over Amazon's Mechanical Turk who were paid $1.00 for their participation.

7.1.2 Procedure and materials

Participants were first introduced to a party scenario with three (randomly generated) characters, here James, Emily, and Robert:

> Yesterday, James, Emily, and Robert went to a big, crazy party that lasted all day. Everyone was playing games and having fun. In this study, you'll read about James, Emily, and Robert's experiences at the party and answer simple questions.

They then proceeded through three blocks of fifteen trials each, for a total of 45 trials. On each block, one of the three characters introduced initially was the speaker. Speaker reliability was manipulated via a short description of the speaker at the beginning of each block. One speaker was reliable (*sober* condition) and two were unreliable (*drunk* and *court* condition). Block order was randomized. The following are examples of speaker descriptions for the block order *sober - drunk - court*.

Sober:

> Flashback to the party: James likes to comment on events he observes. He also likes to keep a clear mind, so he is staying sober. He's having a great time going around observing what people are doing and commenting on what happens.

Drunk:

> Now you know how James experienced the party. Next up is Emily.
>
> Flashback to the party: Emily is pretty drunk. Her speech is slurred. At one point she thinks she sees a flying pig. But she's having a great time going around observing what people are doing. She describes what happens to anyone who will listen.

Table 1: Overview of trial structure in the different blocks (*sober, drunk, court*) of Exp. 4.

	sober	*drunk*	*court*
Event	At the party, Diane threw 15 marbles into the pool.		The prosecutor says: "Diane threw 15 marbles into the pool. What happened next?"
Speaker	James, who interestedly observed what happened, says,	Emily, who drunkenly observed what happened, says,	Robert, who observed what happened but wants to protect his friends, says,
Utterance	"Some of the marbles sank."		
Question	How many marbles sank?		

Court:

> Now you know how Emily experienced the party, too. Next up is Robert.
>
> It's the day after the party. Robert has been asked to appear in court. He was witness to events at the party that resulted in damaged property. The prosecutor asks him questions about the events that transpired. In each case, John saw exactly what happened. But he also wants to protect his friends.

On each trial, participants performed the same task as in Exp. 2b: they saw a sentence introducing an event (e.g., *At the party, Diane threw 15 marbles into the pool*), followed by a sentence about the speaker (e.g., *James, who interestedly observed what happened, says,*), followed by the speaker's utterance (e.g., *"Some of the marbles sank"*). They were then asked to rate on four sliding scales with endpoints labeled "definitely not" and "definitely", how likely they thought 0%, 1-50%, 51-99%, or 100% of the objects (e.g., marbles) exhibited the effect (e.g., sank). The different speaker scenarios made it impossible to maintain an identical trial structure across blocks, but we attempted to minimize the differences in trial structure across blocks, while including material that reminded participants about the speaker's situation. See Table 1 for an overview of trial structure in each block.

In Exp. 4, only half of the items from Exps. 1-3 were used; those 45 that described events that could plausibly have occurred at a day-long party, while nevertheless spanning the entire range of

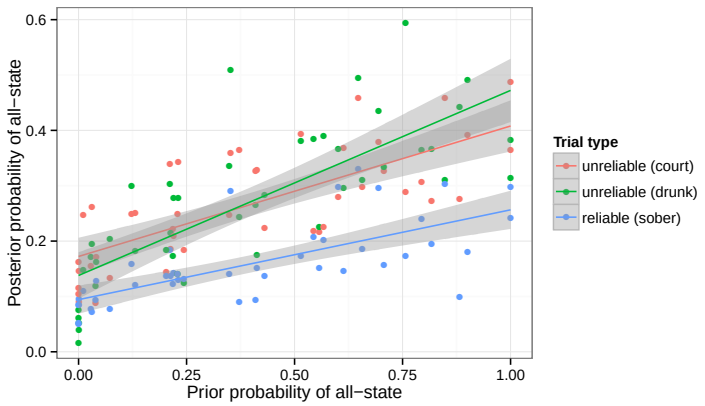

Figure 6: For each item, empirical mean posterior $P(s_{15}|u_{\text{some}})$ against prior $P(s_{15})$ from Exp. 1, separately for reliable (*sober*) and unreliable (*drunk, court*) speakers.

prior probabilities. See Appendix A for the full list of items; the first half of items were used in Exp. 4. Each block contained the same distribution of utterances: two *all* trials, two *none* trials, two short filler trials, two long filler trials, and seven *some* trials. Items were distributed over utterances so that each participant saw each item once, i.e., no participant saw for example the *sinking marbles* item with both the *some* and the *all* utterance.

7.2 Results and discussion

Mean posterior $P(s_{15}|u_{\text{some}})$ ratings for the different speaker reliability conditions are shown in Figure 6. While the reliable (*sober*) speaker condition tracks the results from Exp. 2b very closely, the unreliable speaker conditions (*court, drunk*) display a much greater effect of prior beliefs on the posterior probability of the all-state. This is borne out in a linear mixed effects regression model predicting the posterior all-state probability from fixed effects of prior all-state probability, speaker reliability, trial number (to account for adaptation over the course of the experiment), and their interactions. Speaker reliability was coded as a binary variable (reliable vs. unreliable) and centered. I coded it this way because so far I hadn't really been thinking about trying to tease apart the uninformative vs. untruthful speaker. Don't know if it's worth pursuing the difference between the two. We could also just leave out Figure 5 and describe the pattern. The all-state probability and trial number predictors were likewise centered before entering the analysis. The

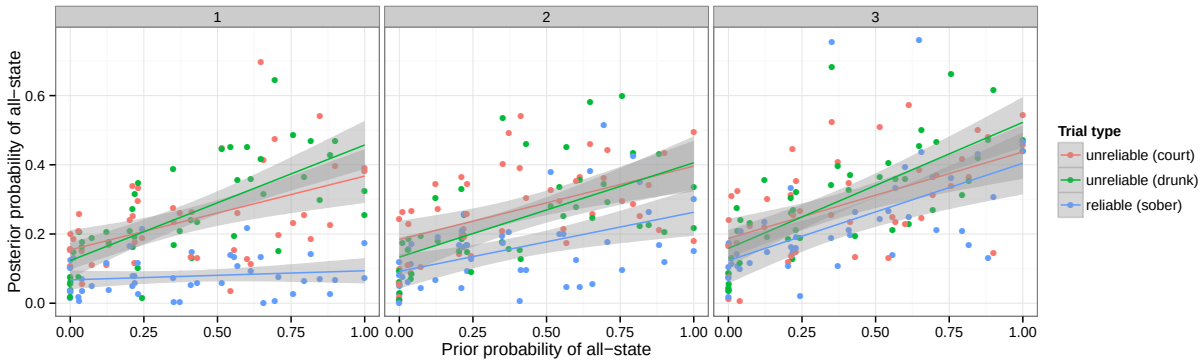

Figure 7: For each item, empirical mean posterior all-state probability against prior all-state probability for reliable (*sober*) and unreliable (*drunk, court*) speakers, by experimental block. Reliable speakers are increasingly treated like unreliable speakers as the experiment progresses.

model included by-participant and by-item random intercepts and slopes for all fixed effects.

There was a main effect of prior all-state probability such that the posterior all-state probability increased with increasing prior all-state probability (β=.26, SE=.03, t=8.5, p<.0001). There was also a main effect of speaker reliability, such that posterior all-state probability was judged higher when the speaker was unreliable (β=.11, SE=.02, t=7.4, p<.0001). In addition, there was an interaction between prior all-state probability and speaker reliability such that the slope of the prior all-state probability effect was shallower for the reliable speaker than for the unreliable speakers (β=.14, SE=.04, t=4, p<.0001). Finally, we also observed adaptation effects: all-state probability was rated as greater as the experiment progressed (β=.003, SE=.0003, t=8.8, p<.0001). We also observed an interaction of prior all-state probability and trial number, such that the prior had a greater effect on interpretation later in the experiment (β=.004, SE=.001, t=3.8, p<.0001). This is visualized in Figure 7: while there is no change in how utterances of unreliable speakers are treated, reliable speakers are treated as more and more unreliable as the experiment progresses.

This is interesting because it suggests that listeners are drawing a higher-level inference about speakers being unreliable in this context (where the context may be this particular party, or this particular experiment), such that participants who had evidence that there were unreliable speakers in this scenario also took the sober speaker less seriously.[10]

[10]Note that in principle the reverse could have happened, that is, evidence for the presence of reliable speakers could have resulted in treatment of the unreliable speakers as more reliable over time. We see no evidence for this.

Taken together, these results suggest that, as predicted, listeners more strongly take into account their prior beliefs when they interpret an utterance produced by an unreliable speaker than one produced by a reliable speaker. In addition, as evidence for unreliability of speakers accumulates, listeners generalize their expectation for unreliability even to a priori reliable speakers, resulting in a greater belief in listeners' priors.

8 Discussion and conclusion

We have shown that listeners' world knowledge, in the form of prior beliefs, enters into the computation of speaker meaning in a systematic but subtle way. The effect of the prior on interpretation was much smaller, and qualitatively different, than predicted by a standard Bayesian model of quantifier interpretation (RSA). This suggests that in certain situations, listeners revise their assumptions about relevant priors as part of the computation of speaker meaning. Indeed, in the cases where the largest deviations from RSA obtained, participants also judged the world to be unusual. Extending RSA with a lifted wonkiness variable that captures precisely whether listeners think the world is unusual, and allows them to back off to a uniform prior (i.e., ignore entirely their previously held beliefs about the world), provided a good fit to the empirical wonkiness judgments and dramatically improved the fit to participants' comprehension data. This model constitutes the first attempt to explicitly model the quantitative effect of world knowledge and its defeasibility on pragmatic utterance interpretation and raises many interesting questions.

In one sense the revision of beliefs in the wRSA listener is standard Bayesian belief updating with respect to a complex prior; however it is not the simple belief update of a flat or hierarchical prior, because the different aspects of prior belief (i.e. $P(w)$ and $P(s|w)$) interact in complex ways with the listener's assumptions about the speaker. As a result, an odd utterance can lead the listener to update their own view of w; this in turn impacts both their own prior over states and what prior they believe the speaker believes they are using—an odd utterance leads the listener to re-evaluate common ground. This is reminiscent of linguistic theories of presupposition accommodation (Lewis, 1979; Stalnaker, 1973, 1998). spell out It will be interesting to further explore the relation of the wRSA approach to presupposition.

Throughout this paper we discussed wonkiness as an attribute of the *world*, yet empirically we elicited wonkiness judgments about the *objects* involved in the events. This raises the question of what exactly listeners are revising their prior beliefs about: objects, events, the speaker's beliefs, or the way the speaker uses language? spell out

Relatedly, we have used a uniform prior distribution as the alternative to consider when the listener believes the world is wonky. One could imagine various more flexible alternatives. For instance, listeners may make minimal adjustments to their prior knowledge, or alternatively, may prefer extreme priors that rationalize the utterance once they have discounted the usual priors.spell out Future research should investigate the options listeners have available when their world knowledge must be revised to accommodate an utterance.

This work also has methodological implications: researchers working in the field of experimental semantics and pragmatics would be well served to take into account the effect of 'odd' items, prior beliefs, and interactions between the two.[11] instead of a footnote, discuss this in more detail, and as a theoretical point about how people should update their beliefs. give it a kleinschmidt spin, i.e. what does this mean for *learning*? short term (local) vs long term (global) belief update. do i infer sth about marbles in general or just the ones in this situation? In particular, if the attempt to design uniform stimuli across conditions yields odd utterances in some conditions, we predict that participants will respond by revising their prior beliefs in ways that can be unpredictable. That is, we expect unpredictable interaction effects between stimuli and conditions. This is likely to inflate or compress potential effects of an experimental manipulation.

Concluding, this work exemplifies the importance and utility of exploring the detailed quantitative predictions of formal models of language understanding. Exploring the prior knowledge effects predicted by RSA led us to understand better the influence of world knowledge and its defeasibility on pragmatic interpretation. Listeners have many resources open to them when confronted with an odd utterance, and re-construing the situation appears to be a favorite.

[11] For an in depth discussion of this issue in syntactic processing, see, e.g., Jaeger (2010); Fine, Jaeger, Farmer, and Qian (2013).

A Items

Items in Exps. 1 - 3 began with a *context sentence* of the form *CHARACTER did Z to X*. In Exps. 2 - 3, the context sentence was followed by a speaker utterance sentence of the form *SPEAKER, who observed what happened, says: UTTERANCE*. On target trials, UTTERANCE was of the form *QUANTIFIER of the X Yed*, where QUANTIFIER was one of *some*, *none*, or *all*. On short filler trials, UTTERANCE was one of *Typical*, *Nothing out of the ordinary*, *As usual*, *Pretty normal*, or *Nothing surprising there*. On long filler trials, UTTERANCE was customized to the particular item in such a way that the utterance addressed an aspect of the situation that was not the number of X that Yed. In the following, we list Z, X, and Y for each of the 90 items as well as its customized filler. We organize items by verb Y, since each verb shared three X and one long filler. CHARACTER and SPEAKER were randomly sampled from a list of 110 names (half male, half female) on each trial. The first 15 items were also used in Exp. 4.

1. Y: melted, X_1: pencils, X_2: crayons, X_3: ice cubes, Z: left in the hot sun
   Filler: It's a beautiful day.

2. Y: stuck to the wall, X_1: baseballs, X_2: cakes, X_3: pieces of gum, Z: threw against a wall
   Filler: What a strange thing to do.

3. Y: burned, X_1: rocks, X_2: books, X_3: matches, Z: threw into a fire
   Filler: I love watching fires.

4. Y: blew away, X_1: backpacks, X_2: hats, X_3: napkins, Z: left on a table on a windy day
   Filler: I just wish the weather was better.

5. Y: ripped, X_1: shoes, X_2: shirts, X_3: books, Z: used as dog toys
   Filler: Doesn't the dog have its own toys?

6. Y: broke, X_1: logs, X_2: boxes, X_3: sunglasses, Z: ran over with a car
   Filler: Why does Lucy always leave her stuff in the driveway?

7. Y: exploded, X_1: candles, X_2: fireworks, X_3: gas tanks, Z: lit
   Filler: Who came up with that idea?

8. Y: dissolved, X_1: carrots, X_2: oreos, X_3: sugar cubes, Z: put in a bucket of water
   Filler: There are people starving in the world.

9. Y: stuck, X_1: beads, X_2: sequins, X_3: stickers, Z: glued to a piece of paper
   Filler: It looks like a zebra.

10. Y: ate the seeds, X_1: dogs, X_2: butterflies, X_3: birds, Z: left seeds out for
    Filler: I wish someone would leave seeds out for me.

11. Y: landed flat, X_1: notebooks, X_2: pancakes, X_3: coins, Z: tossed
    Filler: I love throwing stuff, too.

12. Y: sank, X_1: balloons, X_2: cups, X_3: marbles, Z: threw into a pool
    Filler: It's just fun to throw stuff in the water.

13. Y: fell down, X_1: shelves, X_2: block towers, X_3: card towers, Z: punched
    Filler: Some people just love destruction.

14. Y: rolled, X_1: toy cars, X_2: shopping carts, X_3: wheelchairs, Z: pushed
    Filler: Pushing stuff is so much fun.

15. Y: froze, X_1: bottles of hand soap, X_2: chocolate bars, X_3: berries, Z: put in the freezer
    Filler: That reminds me I need to visit my grandma.

16. Y: flashed, X_1: webcams, X_2: phones, X_3: cameras, Z: took a picture with
    Filler: Everyone with the selfie craze these days.

17. Y: popped, X_1: eggs, X_2: balloons, X_3: bubbles, Z: poked with a pin
    Filler: That requires a lot of concentration.

18. Y: exploded, X_1: CDs, X_2: balls of tin foil, X_3: eggs, Z: heated up in a microwave
    Filler: That's one way of spending your free time.

19. Y: lit up, X_1: cd-players, X_2: computers, X_3: flashlights, Z: pressed the 'on' button on
    Filler: I wish we could just say 'on'.

20. Y: beeped, X_1: houses, X_2: old cars, X_3: new cars, Z: left the lights on in

    Filler: That's not very good for the environment.

21. Y: decomposed, X_1: soda cans, X_2: pinecones, X_3: banana peels, Z: put in a compost pile for a month

    Filler: What a great way to reduce trash.

22. Y: honked, X_1: bicyclists, X_2: bus drivers, X_3: taxi drivers, Z: cut off

    Filler: That looks kind of dangerous.

23. Y: laughed, X_1: lawyers, X_2: comedians, X_3: kids, Z: told a joke to

    Filler: I guess once a jokester, always a jokester.

24. Y: ran out of batteries, X_1: phones, X_2: bike lights, X_3: laptops, Z: left on (and unplugged) all day

    Filler: That would be a pretty useful gadget.

25. Y: had the letter Z in them, X_1: birthday cards, X_2: love notes, X_3: novels, Z: wrote

    Filler: I just don't have the patience to write one of those.

26. Y: were green, X_1: strawberries, X_2: bananas, X_3: clovers, Z: saw

    Filler: I should start growing those myself.

27. Y: got stained, X_1: white tablecloths, X_2: white shirts, X_3: white carpets, Z: spilled red nail polish on

    Filler: Why always the red?

28. Y: reflected the sunlight, X_1: phone screens, X_2: diamonds, X_3: mirrors, Z: placed in the sun

    Filler: Why not just take them inside?

29. Y: rhymed, X_1: poems, X_2: songs, X_3: limericks, Z: wrote

    Filler: It's so pretty.

30. Y: stopped, X_1: bicycles, X_2: motorcycles, X_3: cars, Z: pressed the brakes on

Filler: So many parts need to work for us to not die.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure in mixed-effects models: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255 – 278.

Bergen, L., Goodman, N. D., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the thirty-fourth annual conference of the cognitive science society.*

Degen, J., Franke, M., & Jäger, G. (2013). Cost-Based Pragmatic Inference about Referential Expressions. In *Proceedings of the 35th annual conference of the cognitive science society.*

Fine, A. B., Jaeger, T. F., Farmer, T. F., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8*(10).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, *4*(1), 1–82.

Geurts, B. (2010). *Quantity Implicatures*. Cambridge: Cambridge Univ Press.

Goodman, N. D., & Lassiter, D. (in press). Probabilistic Semantics and Pragmatics: Uncertainty in Language and Thought. *Handbook of Contemporary Semantics, 2nd Edition*.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–84.

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment ? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 169–231.

Kao, J., Wu, J., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(33), 12002–12007.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of salt 23* (pp. 587–610).

Lewis, D. (1969). *Convention. A Philosophical Study.* Harvard University Press.

Lewis, D. (1979). Scorekeeping in a Language Game. *Journal of Philosophical Logic*, *8*(1), 339–359.

Marcus, G. F., & Davis, E. (2013). Psychological Science. (October). doi: 10.1177/0956797613495418

Russell, B. (2012). *Probabilistic Reasoning and the Computation of Scalar Implicatures.* Unpublished doctoral dissertation, Brown University.

Stalnaker, R. (1973). Presuppositions. *Journal of Philosophical Logic*, *2*(4), 447–457.

Stalnaker, R. (1998). On the Representation of Context. *Journal of Logic, Language and Information*, *7*(1), 3–19.