Notes on sinking marbles

Judith Degen

September 11, 2014

Contents

1 Motivation

From Bart Geurts's book "Quantity Implicatures": "The cancellation argument presupposes that preferred scalar inferences will quietly withdraw whenever they happen to be implausible, and this pre- supposition is doubtful, as it is contradicted by the behaviour of bona fide scalar inferences. Let me give a couple of examples to establish this point.

(1) Some of the liberal parliamentarians voted against the bill.

In some democracies, parliamentary fractions tend to vote en bloc. In the Netherlands, for instance, it would be quite unlikely that some but not all liberal parliamentarians voted against a bill. Nevertheless, as far as I can tell, this wouldnt diminish the likelihood that (58) is interpreted as implying that the liberal fraction was divided.

(2) In order to prevent the rinderpest from spreading through the herd, some of the cows were vaccinated.

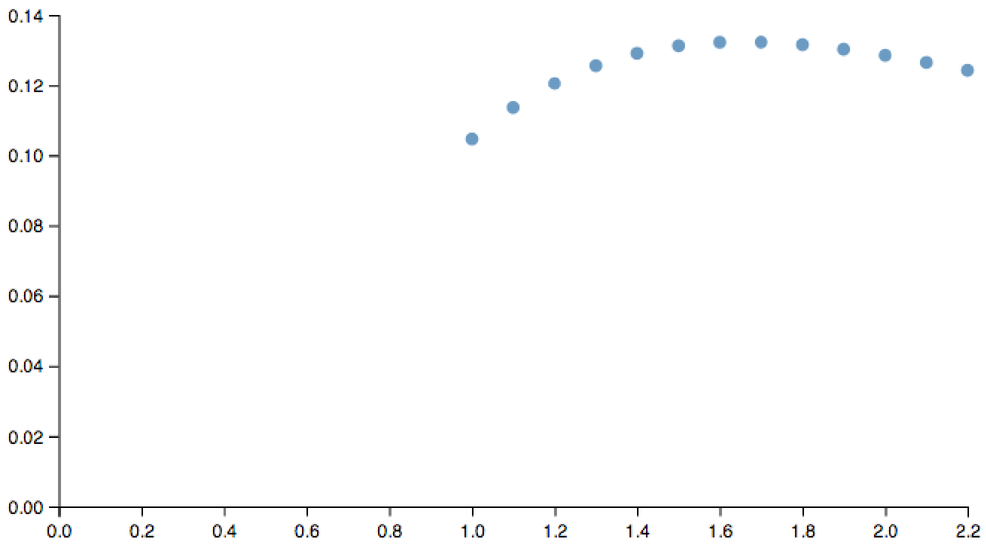


Figure 1: Model predictions for ∀ state after observing *Some of the X sank* with different prior probabilities of ∀ state (i.e. different X types) on the x-axis. Plot generated in play-space with alternative weights for *null; none; some; all* = 20; 1; 1; 1.

Since rinderpest is a highly contagious disease, it would be decidedly odd if only some of the cows were vaccinated, yet that is how we would understand (59).

(3) Cleo threw all her marbles in the swimming pool. Some of them sank to the bottom.

No doubt, it would be very odd if some of Cleos marbles failed to sink, yet according to my intuitions that is precisely what (60) conveys.

What these examples show is that genuine scalar inferences are not so easy to cancel, at all (cf. Geurts 1999a). From a Gricean perspective, this is to be expected: scalar implicatures arise because the speaker goes out of his way to make a statement that is weaker than what he could have said with equal effort, so it stands to reason that it should require special circumstances to suppress a scalar implicature. But in particular, lack of plausibility will generally be insufficient for doing so."

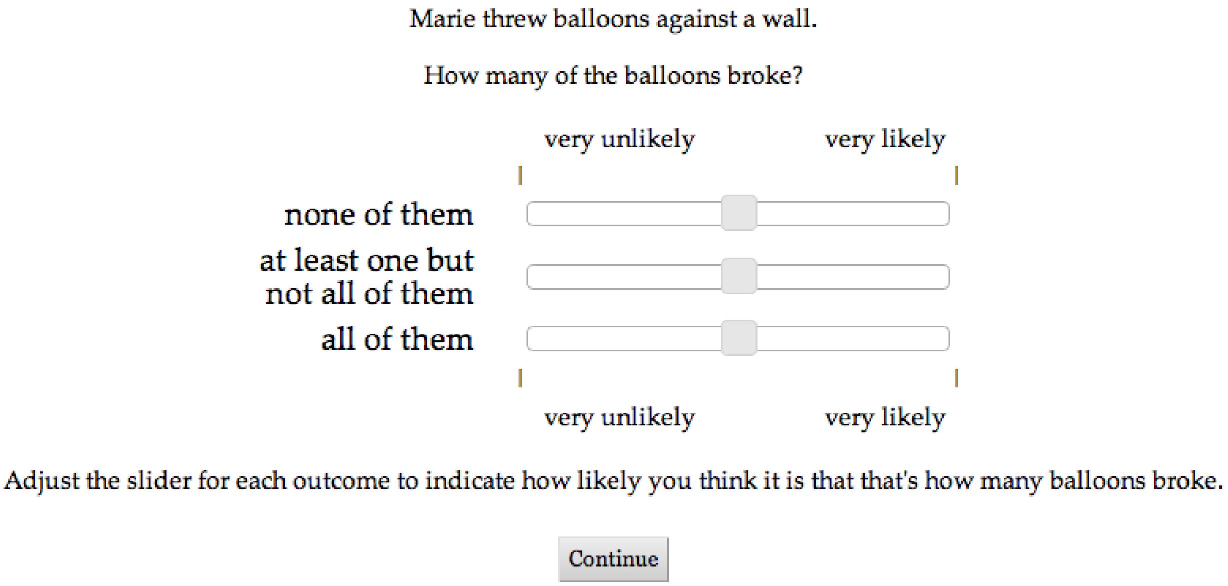
Our questions are: do Bart's intuitions pan out empirically? And if so, how can we model this effect?

General modeling idea: if the prior probability of the ∀ state is high enough and there's a cheap enough alternative utterance –say, silence, or in order to refer to it: *null*– to *some* (crucially, cheaper than *some*) that signifies something like "everything is as usual", then the fact that the speaker went to the (extra) effort of saying *some* instead of *null* means things must really be different than the normal case. But if the normal case is ∀, then by saying *some* the speaker must really mean $\exists \wedge \neg\forall$. That is, there are two counter-acting forces: a) higher prior probability of ∀ state *increases* the posterior probability of ∀ state after observing *some*; b) presence of *null* alternative *decreases* posterior probability of ∀ state after observing *some* (if prior probability of ∀ is very high (how high is an empirical question)).

A preliminary version of the model is here: `http://forestdb.org/models/SFVSome.html` It's got some unnecessary bits leftover from when the model was intended to capture QUD effects. The general shape of the posterior curve for ∀ that it predicts is shown in Figure 1.

2 Experiment 1 - preliminary pilot

The experiment can be found here: `http://stanford.edu/~jdegen/57_sinkingmarbles/alternatives.html`



Marie threw balloons against a wall.

How many of the balloons broke?

Adjust the slider for each outcome to indicate how likely you think it is that that's how many balloons broke.

Figure 2: Example of a display in the prior elicitation phase.

2.1 Method

2.1.1 Participants

30 participants were recruited over Amazon's Mechanical Turk.

2.1.2 Procedure and materials

The experiment consisted of two phases: prior and posterior elicitation. The prior phase was intended to assess participants' prior beliefs about how likely different types of objects are to either *break/sink* prior to observing any additional information. Object types were marbles, feathers, balloons, books, wine glasses, iPhones. See Figure 2 for an example of a prior trial (including slider dependent measure).

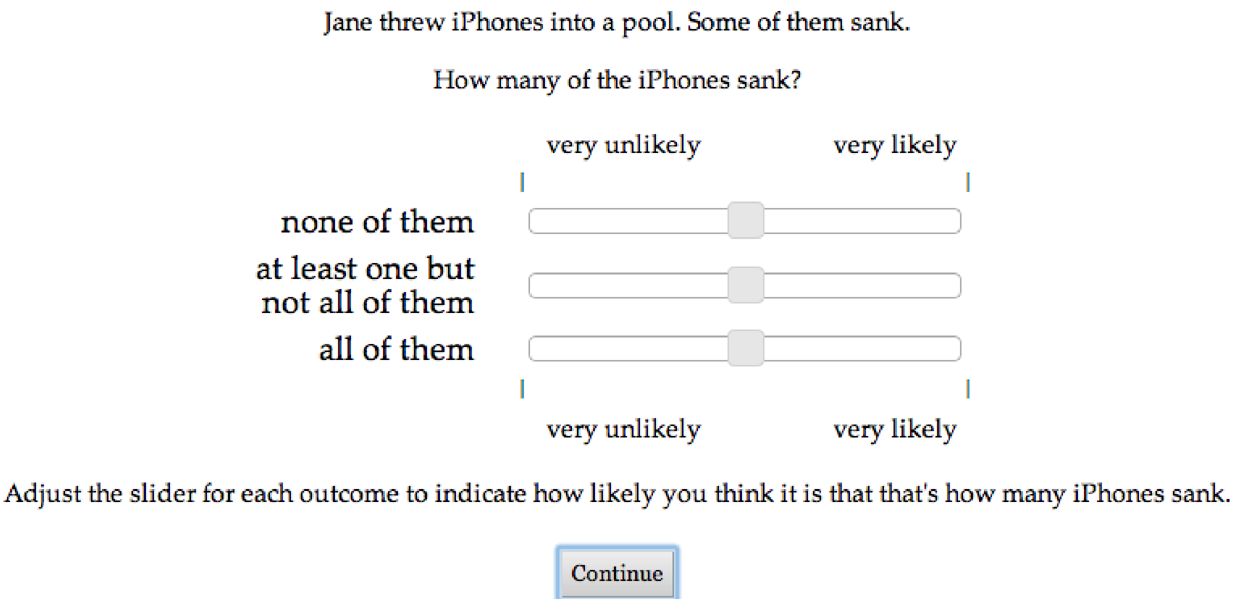
In the posterior elicitation phase, additional information about how many of the objects sank/broke was given by adding an additional sentence: either *All* or *Some of the X broke/sank*. See Figure 3 for an example of a posterior trial.

Each phase of the experiment consisted of 12 trials: each of the six object types occurred with each verb (break/sink) in each phase. In the posterior phase, an object occurred once with *all* and once with *some*, but *all/some* were randomly paired with *break/sink*. That is, for each combination of verb/object we obtained 30 prior probabilities and for each verb/object/quantifier combination we obtained 15 posterior probabilities.

2.2 Results

Slider values were normalized for each trial to yield a probability distribution over states $\neg\exists$, $\exists \wedge \neg\forall$, and $\forall$ for each verb/object combination in the prior phase and each verb/object/quantifier combination in the posterior phase. Figure 4 shows the mean prior probabilities of the $\forall$ state for each object/verb combination.

The interesting question was whether we would see something like the shape in Figure 1. The empirical results are shown in Figure 5 (unaggregated) and Figure 6.



Jane threw iPhones into a pool. Some of them sank.

How many of the iPhones sank?

none of them

at least one but not all of them

all of them

Adjust the slider for each outcome to indicate how likely you think it is that that's how many iPhones sank.

Figure 3: Example of a display in the posterior elicitation phase.

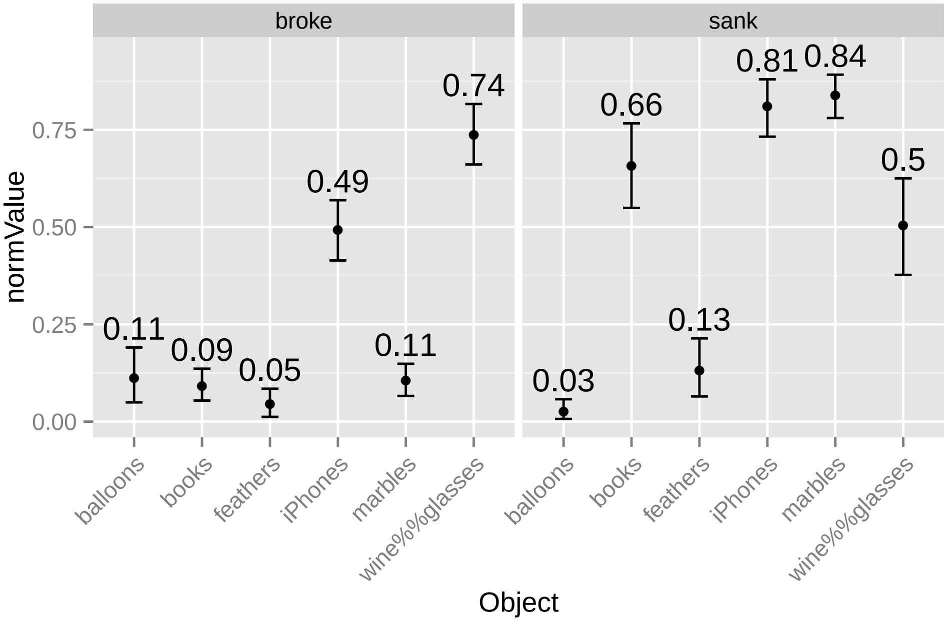


Figure 4: Mean prior probability of all objects breaking (left) or sinking (right) by object type.

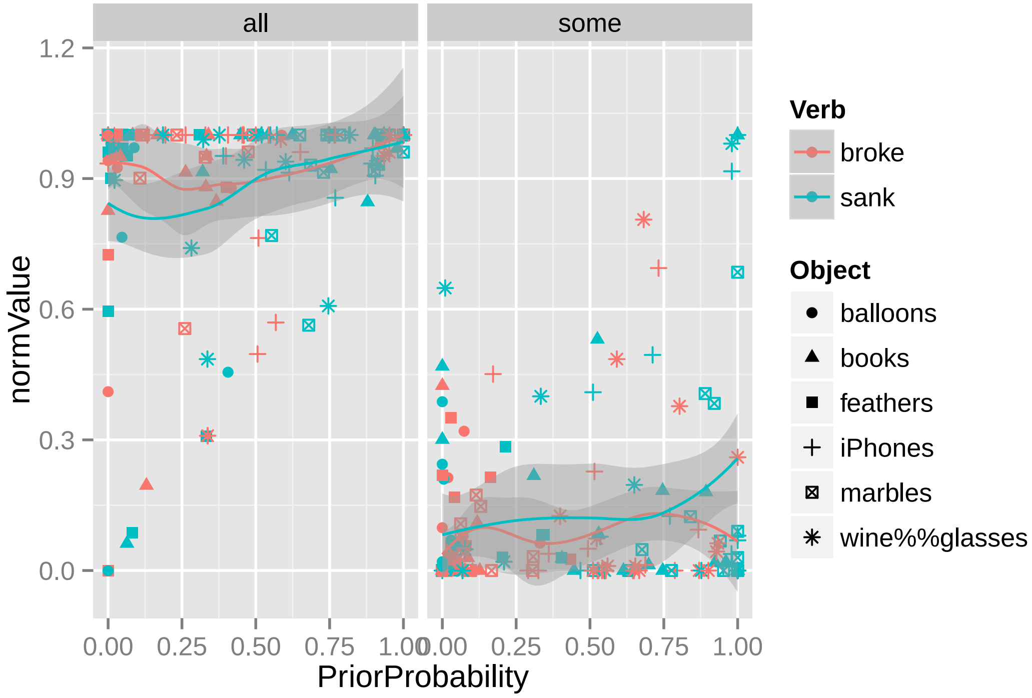


Figure 5: Posterior probability of each object breaking/sinking after being informed that *all* (left) or *some* (right) of them broke, by unaggregated (by-subject) prior probability of all of them breaking/sinking.

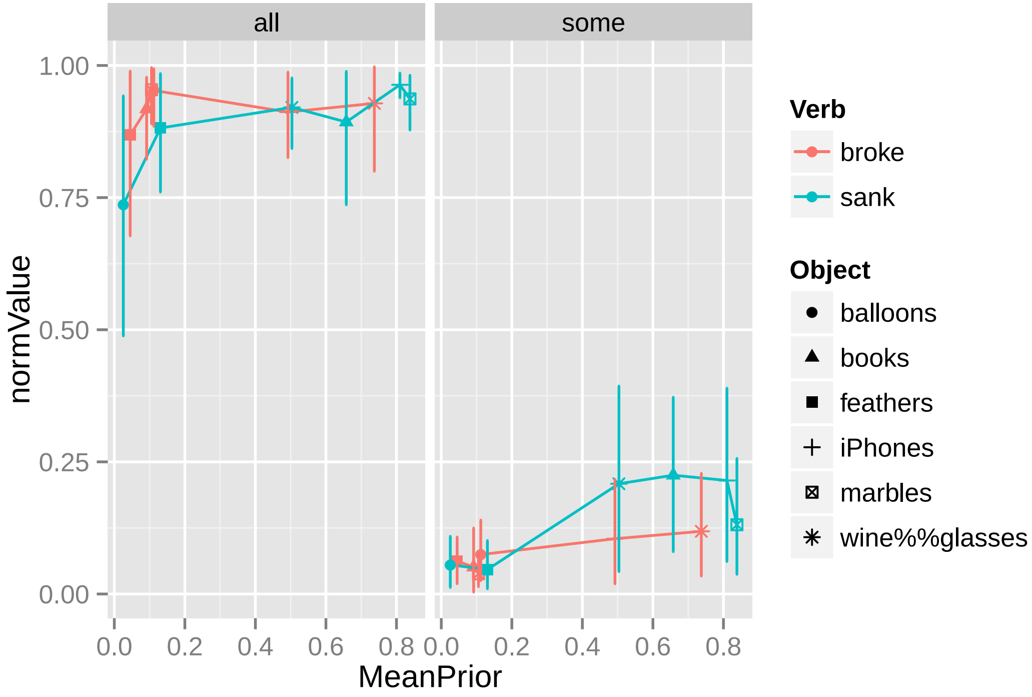
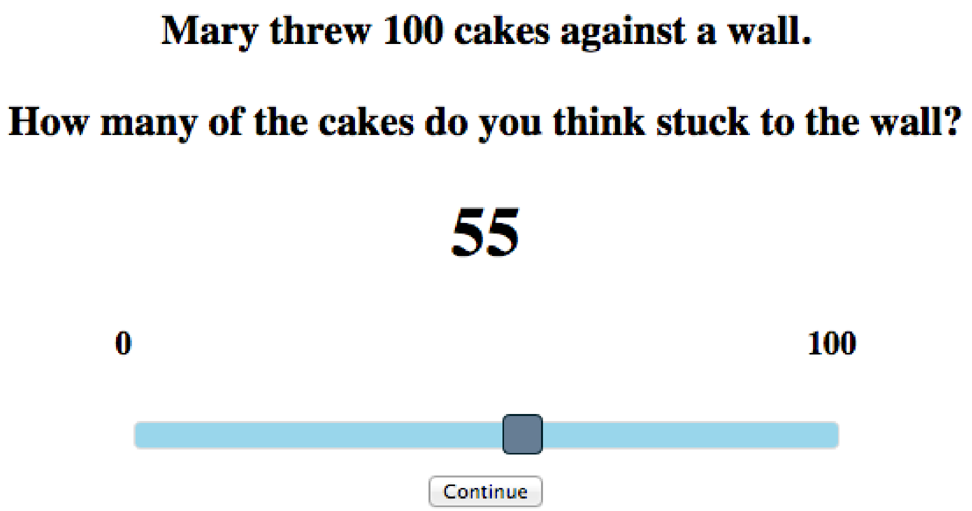


Figure 6: Posterior probability of each object breaking/sinking after being informed that *all* (left) or *some* (right) of them broke, by aggregated prior probability of all of them breaking/sinking. Error bars indicate bootstrapped 95% confidence intervals.


Mary threw 100 cakes against a wall.

How many of the cakes do you think stuck to the wall?

Figure 7: Display in prior elicitation experiment 2.

2.3 Discussion

The unaggregated data look super noisy, the aggregated ones seem to at least sort of show the right shape in the *sink* but not the *break* condition, but the error bars are huge. That it doesn't show up even a little bit in the *break* condition might be because none of the object types has a prior probability of breaking that's high enough (the highest we got was .74 for wine glasses).

Generally we have the problem so far that we're over-sampling the low- and under-sampling the high-prior-probability space. That is, we need more object types with a better spread, and preferably clustered around the higher range.

Another thing: one of the reasons the results might be so noisy is that the *all/some* information is not presented as part of a speaker's utterance, but as part of the instructions on each trial.

So. . .

3 Experiment 2 - prior elicitation

The goal of this experiment was to pilot a series of items and make sure there was a large range of prior probabilities for different events and especially, that we got many samples from the highest region, i.e. probability>.8. The experiment can be found here: `https://web.stanford.edu/~erindb/sinking-marbles/experiments/sinking-marbles-prior/sinking-marbles-prior.html`

3.1 Method

3.1.1 Participants

60 participants were recruited over Amazon's Mechanical Turk and paid $0.50 each.

3.1.2 Procedure and materials

On each trial, participants saw a display as in Figure 7, which consisted of a sentence frame "SPEAKER VERB-ed 100 OBJECTs EXTRA", e.g. *Mary threw 100 cakes against a wall.* They were then asked to provide an estimate for how many OBJECTs displayed a certain EFFECT, e.g. sticking to the wall, by adjusting a slider.

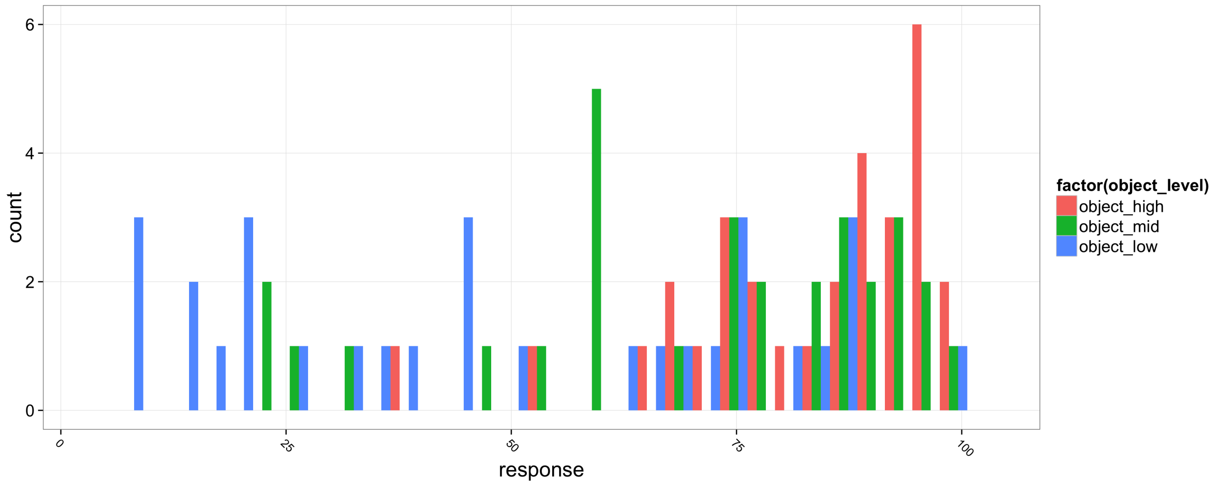


Figure 8: Histogram of elicited mean priors.

Three items were created for each VERB-EXTRA combination, corresponding to three different OBJECT types. OBJECT types were chosen in such a way as to result in low, medium, and high prior probabilities of displaying the EFFECT. Each participant saw each VERB-EXTRA combination with only one OBJECT type. Each participant saw a total of 30 trials.

3.2 Results

See Figures 8 and 9.

3.3 Discussion

This is a nice range and especially covers the highest end of the spectrum. Some of the items displayed some idiosyncrasies – issues of EFFECT being a state/inherent property rather than a point outcome (accomplishment or whatever the linguists call it). And differences in how long the agent needs to be present to observe the EFFECT. Does it matter? Let's assume for the time being it doesn't.

4 Experiment 3 - sinking marbles, version 1

The goal of this experiment was to test for the predicted 'sinking marbles' effect. The experiment can be found here: `https://web.stanford.edu/~erindb/sinking-marbles/experiments/sinking-marbles/sinking-marbles-prior.html`

4.1 Method

4.1.1 Participants

120 participants were recruited over Amazon's Mechanical Turk and paid $??? each.

4.1.2 Procedure and materials

On each trial, participants saw a display as in Figure 10, which consisted of a sentence frame "PERSON VERB-ed N OBJECTs EXTRA", e.g. *Mary threw 12 cakes against a wall*. They were also told that a SPEAKER, who observed

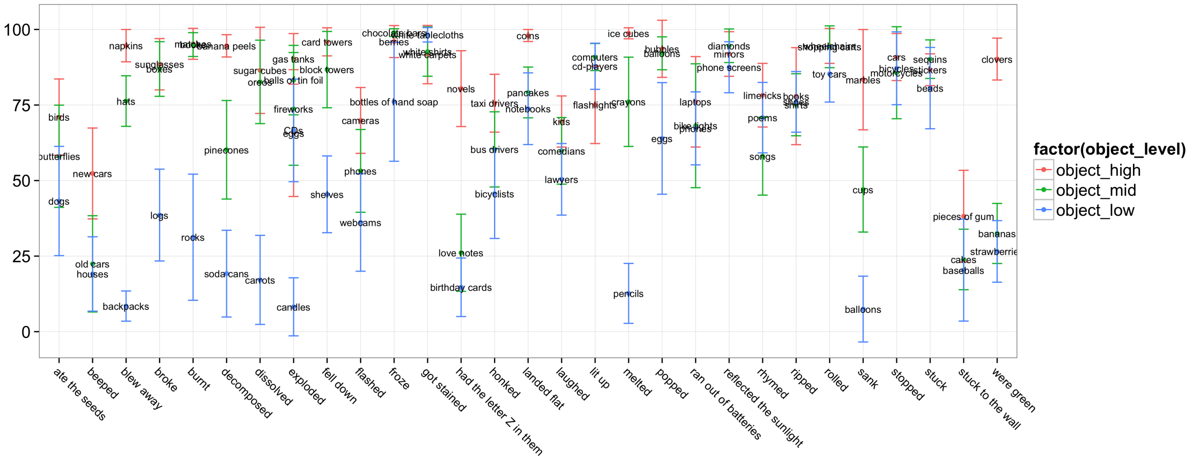


Figure 9: Elicited mean priors by object type.

what happened, produced an utterance. The utterance was always of the form "QUANTIFIER of the OBJECTs EFFECT.", e.g. *All of the cakes stuck to the wall*. They were then asked to provide an estimate for how many OBJECTs displayed the EFFECT, e.g. sticking to the wall, by adjusting a slider for each of 4 options: 0%, 1-50%, 51-99%, or 100%.

The same items were used as in Exp. 2. Each participant saw each VERB-EXTRA combination with only one OBJECT type and one QUANTIFIER. QUANTIFIER could be one of *all, most, some* or *none*, and was randomized. N varied randomly from 4 to 15. PERSON and SPEAKER also varied randomly. Each participant saw a total of 30 trials.

4.2 Results

We were interested in testing the model prediction that when observing an utterance with *some* a) the probability of the all-state obtaining should increase with increasing prior probability, and b) the probability of the all-state obtaining should flatten out/decrease again for the highest-prior events because of the availability of saying something cheap or nothing that signals "everything is as usual".

First, sliders were normalized to give a proper probability distribution over proportions (0, .01-.5, .51-.99, 1) on every trial. Figure 11 shows the smoothed mean posterior probabilities as a function of prior probability, separately for each quantifier and proportion.

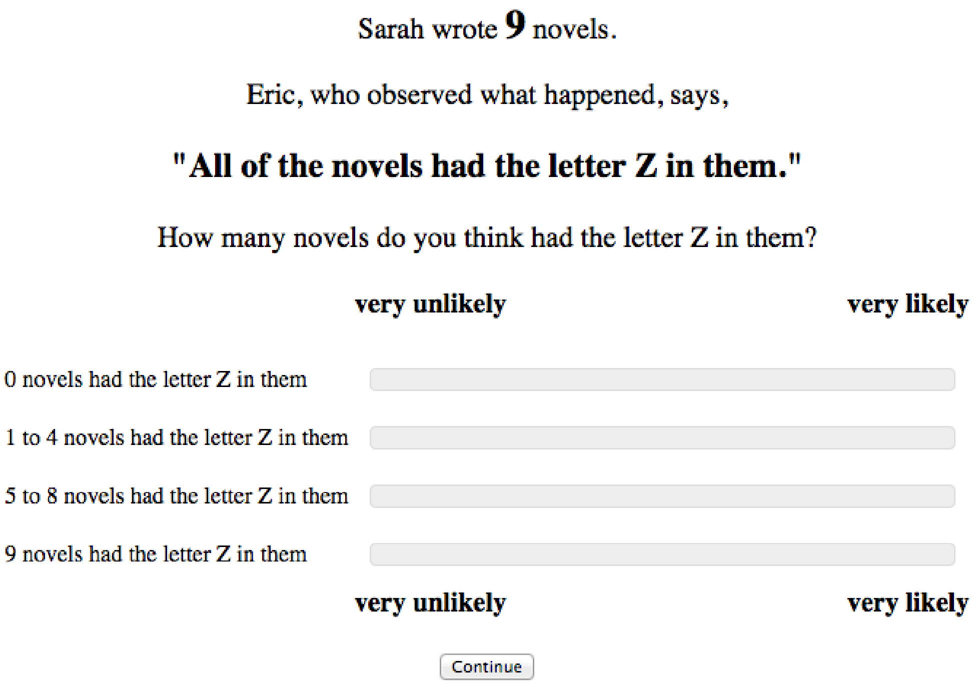
4.3 Discussion

The interesting column is the last one (100%), which shows the mean by-item probability of the all-state obtaining. First thing to notice: effect of prior is super small. There's definitely a positive slope for both *some* and *most*, but it's a tiny effect. As for the flattening out effect - it seems to be there for *most*, but not for *some*. Splitting results up by quarter (which really shouldn't be done because there's not enough data...) indicates that whatever effect there is, it's stronger in the first quarter than in the others, where things are just flat.

So...

5 To do

run a version with the following features:



Sarah wrote **9** novels.

Eric, who observed what happened, says,

"All of the novels had the letter Z in them."

How many novels do you think had the letter Z in them?

0 novels had the letter Z in them

1 to 4 novels had the letter Z in them

5 to 8 novels had the letter Z in them

9 novels had the letter Z in them

Figure 10: Display in sinking marbles experiment 3.

- no *most* (because that might be eating up some of the dynamic range we'd otherwise get for *some*)
- two types of fillers to mimic a null alternative (because people may be restricting the set of alternatives to just sentences of the form "Q of the O VERB-ed", but that defeats the purpose of assuming they're entertaining a null alternative): short fillers of the sort "Typical/normal/nothing out of the ordinary" and other generic things that signal that *everything is as usual/expected*. And long fillers that basically shift the QUD away from "How many OBJECTs VERBed?", thus signaling that it's not the quantity of the OBJECTs that VERBed that's interesting –that's as usual– but rather something else.

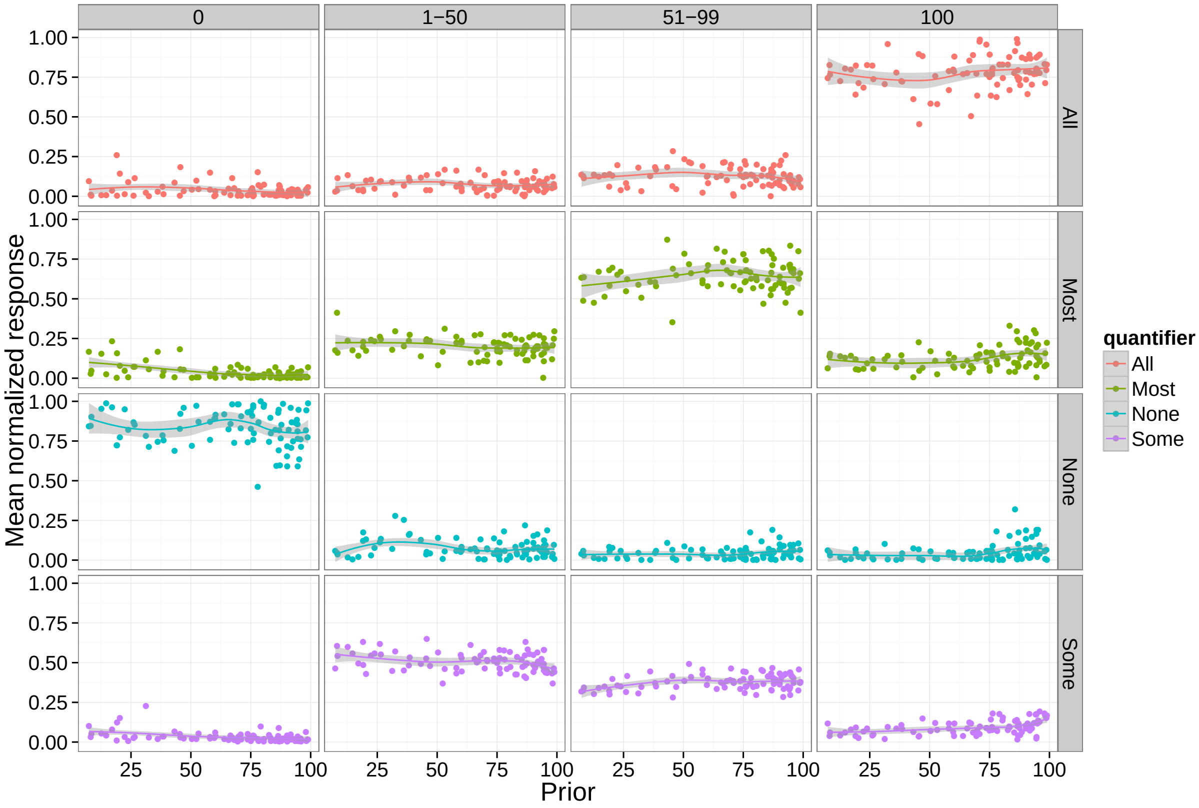


Figure 11: Smoothed mean posterior probabilities as a function of prior probability, separately for each quantifier and proportion.