Lost your marbles? The puzzle of dependent measures in experimental pragmatics

Judith Degen (jdegen@stanford.edu), **Noah D. Goodman** (ngoodman@stanford.edu)
Department of Psychology, Stanford University
Stanford, CA USA 94350

Abstract

A rarely discussed but important issue in research on pragmatic inference is the choice of dependent measure for estimating the robustness of pragmatic inferences and their sensitivity to contextual manipulations. Here we present the results from three studies exploring the effect of contextual manipulations on scalar implicature. In all three studies we manipulate the salient question under discussion and the perceptual availability of relevant set sizes. The studies differ only in the dependent measure used: Exp. 1 uses truth judgements, Exp. 2 uses word probability ratings, and Exp. 3 uses a direct measure of sentence interpretation. We argue that the first two are effectively measures of production, and find they are sensitive to our contextual manipulations. In contrast the interpretation measure shows no effect of context. We argue that this methodologically troubling finding can be understood and predicted by using the framework of probabilistic pragmatics.

Keywords: pragmatics; psycholinguistics; scalar implicature; QUD; methodology

Introduction

Context affects language understanding in complex and profound ways. Relatively small changes in the context of an utterance can radically change the interpreted meaning—or fail to entirely. For instance, take a *scalar implicature*: "some of the candies in that box contained nuts" will often be pragmatically strengthened to "and not all of them did". However, finding out that the speaker ate only two candies (thus lacked knowledge of the stronger situation) or has a nut allergy (thus cares only if *any* candies contain nuts) can decrease the strength of this implicature (i.e., make it more plausible that all the candies had nuts).

Pragmatic judgments, e.g., about scalar implicatures, are notoriously volatile, responding not only to minor aspects of the cover story, but also to the dependent measure used to probe linguistic intuitions (Geurts & Pouscoulous, 2009; Zondervan, 2010). Contradictory results from different dependent measures are particularly troubling to progress in the empirical study of pragmatics. Here we systematically explore three dependent measures of scalar implicature, varying the context, aiming to relate them to each other and to recent formal models of pragmatics. We offer a potential explanation for why some measures are more sensitive to context manipulation than others.

Most experimental studies on scalar implicature have used either sentence verification or metalinguistic judgment paradigms to probe participants' interpretation of utterances containing scalar items (Bott & Noveck, 2004; Geurts & Pouscoulous, 2009; Geurts, 2010; Zondervan, 2010; Degen & Tanenhaus, to appear). In sentence verification studies, participants are often shown a visual display or asked to read a story, thus establishing the facts about the world. They are then asked to provide a binary judgment about the truth of an utterance containing a scalar item, such as *Some of the B's are in the box on the left*. In metalinguistic judgment paradigms, participants are shown an utterance like *Some of the B's are in the box on the left* and are then asked explicitly whether it follows that not all of the B's are in the box on the left (Geurts & Pouscoulous, 2009). Yet neither of these dependent measures seems to directly measure the natural modes of language: production and interpretation. In everyday life people constantly produce and interpret language, but much more rarely adjudicate the truth of a sentence. Occasionally researchers have used measures that more directly evaluate interpretation (e.g. Goodman & Stuhlmüller, 2013) but some evidence—as well as much anecdotal experience—suggests that these measures may be less sensitive and more unstable.

Methodologically, it is unsettling that implicature rates differ depending on the dependent measure, and that different dependent measures are sensitive to contextual manipulations to different degrees. How is one to choose a dependent measure that adequately reflects the underlying contextual effect, if there is one? We will not be able to answer this question fully here. Rather, we will present results from three experiments that provide the first step towards a full investigation into the sensitivity of different dependent measures to manipulations of context and the underlying pragmatic inference process. We manipulate the dependent measure that is used to evaluate listeners' sensitivity to two contextual cues when they interpret utterances containing scalar items: a) the implicit contextual Question Under Discussion (QUD, Roberts, 2004) that interlocutors are trying to address; and b) the size of the set that the quantifier *some* is being used to describe. The experiments we present here are identical with the exception of participants' task.

Exp. 1 uses *sentence verification*, the most widely employed measure in the literature. We will speculate that sentence verification is closely related to production; to evaluate this idea we perform Exp. 2, which uses *word probability ratings* as a more direct test of listeners' expectations about speakers' production. Finally, Exp. 3 uses *sentence interpretation*—a measure that constitutes arguably the most direct and natural test of understanding by allowing participants to distribute confidence over different potential states of the world in response to an observed utterance.

We delay discussion of the differences between these measures to the General Discussion, where we provide an argument that the differences in the results of Exps. 1 and 2 vs. those of Exp. 3 are expected under a formal, probabilistic model of pragmatics.

Experiment 1 - sentence verification

The two contextual features we investigate are the implicit question that interlocutors are trying to address (the QUD) and the total size of the set of objects, known by the speaker, that is contextually available. We elaborate briefly on each.

Previous studies have found, using sentence verification, that a contextually evoked QUD like *Did I get all of the gumballs?* yields higher implicature rates than a QUD like *Did I get any of the gumballs?* when participants are presented with a scenario in which they got all of the gumballs but are told *You got some of the gumballs* (Degen, 2013). A similar effect was obtained by Zondervan (2010). This effect is presumably due to the former QUD making the stronger scalar alternative *You got all of the gumballs* more relevant than the latter QUD does, which in turn has the effect of listeners inferring that the speaker must have really meant that they did not get all of the gumballs if she didn't use the relevant *all* alternative. In the following we will refer to a QUD that makes the utterance with *all* relevant as the *all?* QUD and a QUD that make only the lower bound (*at least one*) relevant as the *any?* QUD.

The second contextual feature of interest is the total number of objects in the domain. Previous studies have found that *some* is more natural to describe sets of sizes that are not subitizable (Degen & Tanenhaus, to appear; van Tiel, 2013). This is presumably due to the increased effort it takes speakers to establish exact set size for larger sets—using a vague quantifier to describe a larger set allows speakers to hold the floor without saying untrue things. This predicts that listeners should expect speakers to be less likely to use *some* to refer to all the objects in a set when that set is subitizable (e.g. of cardinality 4) than when it is not (e.g. of cardinality 16).

Exp. 1 tests the effect of QUD and set size using sentence verification, the most widely used measure of scalar implicature derivation in the literature (Bott & Noveck, 2004; Degen & Tanenhaus, to appear; Zondervan, 2010; Geurts & Pouscoulous, 2009). Participants are given the facts about the world and a speaker's utterance and are asked to judge the truth of the utterance.

Method

Participants We recruited 48 participants over Amazon's crowd-sourcing platform Mechanical Turk.

Procedure and materials Participants read a brief two-paragraph story about a character (henceforth, the speaker) who lost her marbles in shoe boxes when her nephew came over to play. The story introduced a number of marbles (either 4 or 16) that the speaker owned and evoked an implicit QUD that made it relevant either that the speaker find *all* of the lost marbles (*all?* condition) vs. *at least one* of them (*any?* condition). An example of an *all?* context is the following:

> Ann is really into collecting marbles. Recently, her friends gave her a special edition of 16 marbles, which she loves. Yesterday, her five-year-old nephew came to visit and found her set of marbles in a drawer. He also found some shoe boxes. He played with the marbles for a long time and moved them from one box to another until they were all hidden and he didn't remember where he put them.
>
> When Ann later entered the room, she saw that all the marbles were gone and there was a pile of shoe boxes on the floor. She was upset and complained bitterly to her husband. She was determined to find every last one of her marbles. She started opening one box after another, looking for marbles.

In contrast, contexts like the following were employed to evoke the *any?* question:

> Ann's five-year-old nephew loves playing with marbles. For when he comes to visit, Ann keeps a set of 16 marbles in a drawer. Yesterday, he came to visit and found her marbles in the drawer. He also found some shoe boxes. He played with the marbles for a long time and moved them from one box to another until they were all hidden and he didn't remember where he put them.
>
> When Ann later entered the room, she saw that all the marbles were gone and there was a pile of shoe boxes on the floor. Her nephew was upset because he wanted a marble to play with. He started to cry and Ann's husband tried to console him while Ann started opening one box after another, looking for marbles.

In order to ensure that participants paid attention to the story, they were then asked two questions: *How many marbles are there in Ann's set?* and *When will Ann be satisfied?* They were only allowed to proceed once they correctly answered the questions. The correct answers were either *4* or *16* (to the first question), and *if she finds all of the marbles* or *if she finds at least one of the marbles* (to the second question), respectively. They were then shown a target display containing a box with the complete set of marbles and were told that this was the box that the speaker found. They were also shown the speaker's utterance *I found some of the marbles* and were asked to judge whether the statement was true by clicking either a 'Yes' or a 'No' button.[1] If the speaker is interpreted as saying that she did not find all of the marbles, reflecting an implicature was drawn, participants should respond 'No'. If instead they interpret the utterance literally, they should respond 'Yes'.

QUD and total number of marbles were between-participant manipulations. Speaker name and gender, gender of the visiting relative, and marble color were randomized. Marble location was determined by adding noise to fixed initial positions within the box.

Prediction If listeners are sensitive to the QUD, participants should be more likely to respond 'No', reflecting the implicature, when the QUD is *all?* than when it is *any?*.

[1] The same experiment was run where the task was not to judge truth, but rather whether participants *agreed* or *disagreed* with the speaker. The results were qualitatively the same.

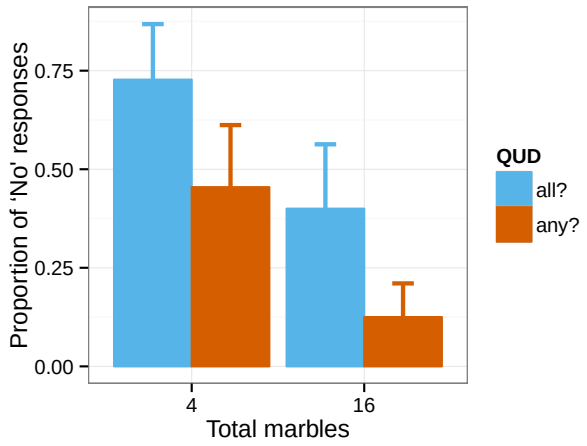

Figure 1: Proportion of pragmatic 'No' responses in Experiment 2 by QUD and total number of contextual marbles.

If listeners further expect speakers to use *all* when it is easy to verify that all the marbles are present (i.e., when the set is subitizable) but less so when the speaker would have to invest effort to determine set size (i.e., when the set is not subitizable), participants should be more likely to respond 'No' for the small than for the big set.

Results and discussion

Proportion of pragmatic 'No' responses are shown in Figure 1. In a logistic regression predicting 'Yes' responses from centered predictors of QUD, SETSIZE, and their interactions, there were main effects of QUD and SETSIZE such that participants were more likely to respond pragmatically ('No') when the question was *all?* (β = 1.37, *SE* = 0.68, $p < .05$) and when set size was small (β = -1.6, *SE* = 0.68, $p < .02$).

This suggests that listeners, despite being asked about the *truth* of the utterance, display pragmatic effects: when the stronger alternative *I found all of the marbles* matters to the contextual QUD they are more likely to reject the *some* statement than when it is not. In addition, the main effect of SETSIZE suggests that listeners took into account the uncertainty that the speaker had about the actual size of the larger set and were aware of the presumably extra counting effort that the speaker would have had to invest in order to verify the precise size of the set.

Using sentence verification, we thus uncovered two contextual effects on scalar inference. If participants had been judging literal truth, we would not have expected any context effects. Why do we nevertheless observe these effects? One possibility is that participants are reinterpreting the task to answer a question like *Could a speaker have said this if he had intended to communicate this particular state of the world?*, suggesting that sentence verification is a measure that probes listeners' intuitions about *production* rather than measuring literal truth or comprehension directly. If this is right, a more direct measure of production should replicate the contextual effects.

Experiment 2 - word probability rating

Exp. 2 used a more direct measure of participants' expectations about production.

Method

Participants We recruited 52 participants over Mechanical Turk.

Materials and procedure Exp. 2 differed from Exp. 1 only in the speaker's utterance and participants' task. Again, they saw the box with the complete set of marbles and were told that that was the box the speaker found, but rather than the speaker's utterance being *I found some of the marbles*, it was *I found _____ of the marbles*. Participants were then shown four utterance alternatives—*some*, *all*, *none*, and the number term *four* or *sixteen*, respectively—and asked to adjust a slider for each word to indicate how likely they thought it was that the speaker used that word. The endpoints of the sliders were marked as *very unlikely* and *very likely*.

Prediction If speakers are less likely to use *some* and more likely to use *all* when the question is *all?* and listeners are sensitive to this, higher slider values for *some* should be obtained when the question is *all?* than when it is *any?*.

If listeners further expect speakers to not go to the effort of counting the big set in order to be able to use the stronger *all* alternative, higher slider values for *some* should be obtained for large than small sets.

Results and discussion

Participants' ratings were normalized such that for each participant, their slider values summed to 1. Mean ratings for each utterance are shown in Figure 2. In a linear regression predicting ratings for *some* from centered predictors for QUD, SETSIZE, and their interactions, the main effects of QUD and SETSIZE and the lack of interaction found in Exp. 2 replicated. Participants were more likely to indicate lower probability of use for *some* when the question was *all?* (β = .16, *SE* = .06, *t* = 2.84, $p < .007$) and when set size was small (β = -.16, *SE* = .06, *t* = 2.81, $p < .008$). The results were qualitatively the same when the analysis was performed on unnormalized slider values. These results mimic the results from Exp. 1 that used sentence verification.

Interestingly, these pragmatic effects are also reflected in participants' expectations of use for *all*: ratings were higher for *all* when set size was small (β = .14, *SE* = .07, *t* = 2.18, $p < .04$) and marginally higher when the question was *all?* (β = -.12, *SE* = .07, *t* = -1.76, $p < .04$), suggesting that even the use of *all*, a quantifier that is typically treated as having a fixed semantics that does not allow for pragmatic slack, is context-dependent. There were no pragmatic effects on expectation of use for number terms and *none*.

Experiment 3 - sentence interpretation

Exp. 3 used a measure of utterance *comprehension* that is arguably more natural than sentence verification: rather than

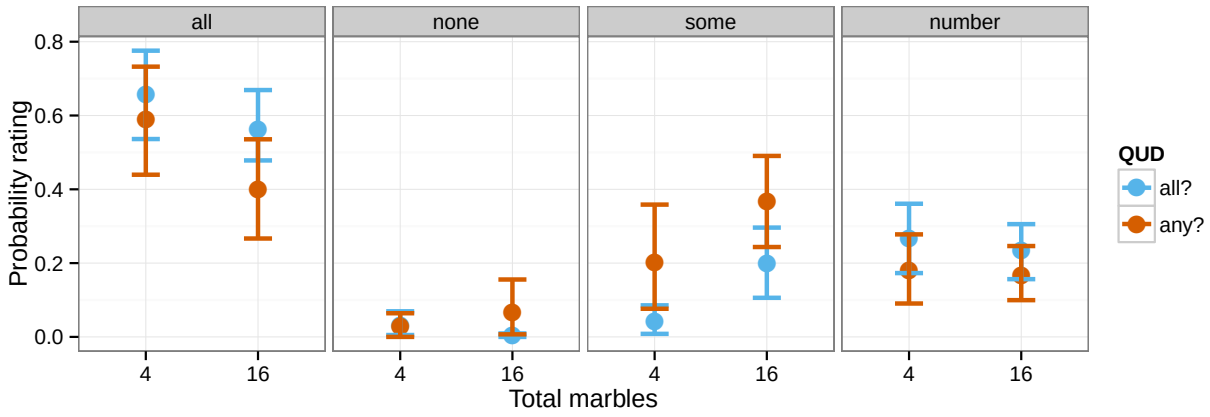

Figure 2: Normalized probability ratings for the four utterance options, by QUD and total number of contextual marbles. Error bars indicate bootstrapped 95% confidence intervals.

being given the facts about the world and an utterance describing those facts, we present participants with the more naturally occurring situation of observing an utterance that a speaker produced—*I found some of the marbles*—and having to infer what the world is that the speaker is trying to communicate.

Method

Participants We recruited 48 participants over Mechanical Turk.

Materials and procedure Exp. 3 was identical to Exps. 1 and 2 with the exception of the target display and participants' task. Rather than seeing just the box with the complete set, participants instead saw a target display as in Figure 3, which contained the speaker's utterance—*I found some of the marbles*—and five hypothetical boxes of marbles that the speaker could have found. Boxes contained (in order) 0, 25, 50, 75, and 100% of the total number of marbles introduced in the context.

Participants' task was to adjust a slider for each box to indicate how likely they thought it was that the speaker had found that box. Once all the sliders were adjusted they proceeded to the post-experiment questionnaire.

Prediction The slider adjustment task allows us to obtain a probability distribution over situations that participants are taking the speaker to convey. If the QUD modulates this distribution, we should observe that when the QUD is *all?*, the distribution is shifted to the left, compared to when the QUD is *any?*. This should become most apparent in the ratings for the box with the complete set: ratings should be higher when the question is *any?* than when it is *all?*.

For the set size manipulation, participants should be more likely to expect *some* to be used with the complete set when set size is not subitizable and determining whether all of the marbles were found would require counting. This should be reflected in higher ratings for the complete set when the total number of marbles is large (not subitizable) than when it is small (subitizable).

It is plausible to expect that effects of pragmatics would be largest with this dependent measure, certainly compared to truth judgement, since it is for comprehension that pragmatic inference is most important. However previous research has indicated that interpretation is a particularly fragile measure.

Results and discussion

Participants' ratings were normalized such that for each participant, slider values summed to 1. Mean ratings for each box are shown in Figure 3. Mean ratings for the complete set (containing 4 or 16 marbles, respectively) were close to floor and ranged from 0.09 to 0.13. In a linear regression predicting ratings for the complete set from centered predictors for whether the QUD was *any?*, whether the total set size was small, and their interactions, no effects reached significance (all $ps > .39$). Results were qualitatively the same when the analysis was performed on unnormalized slider values.

We thus find no evidence for effects of QUD and set size on the probability of generating a scalar implicature with the sentence interpretation measure, despite this measure being the most direct measure of comprehension. We discuss this surprising result and the asymmetry with Exps. 1 and 2 in the following.

General discussion

The empirical pattern of obtained results is the following: an effect of QUD and set size was observed when using the sentence verification and word probability rating measures (in the same and predicted direction, Exps. 1 and 2), but not when using the sentence interpretation measure (Exp. 3). It is surprising that sentence interpretation, which is the task that most directly mimics the task that listeners are confronted with every day—inferring the state of the world that a speaker is most likely intending to communicate—is the one that does not show sensitivity to context manipulations; while at the same time the less natural measures of interpretation—judging whether a sentence can be uttered in a given context

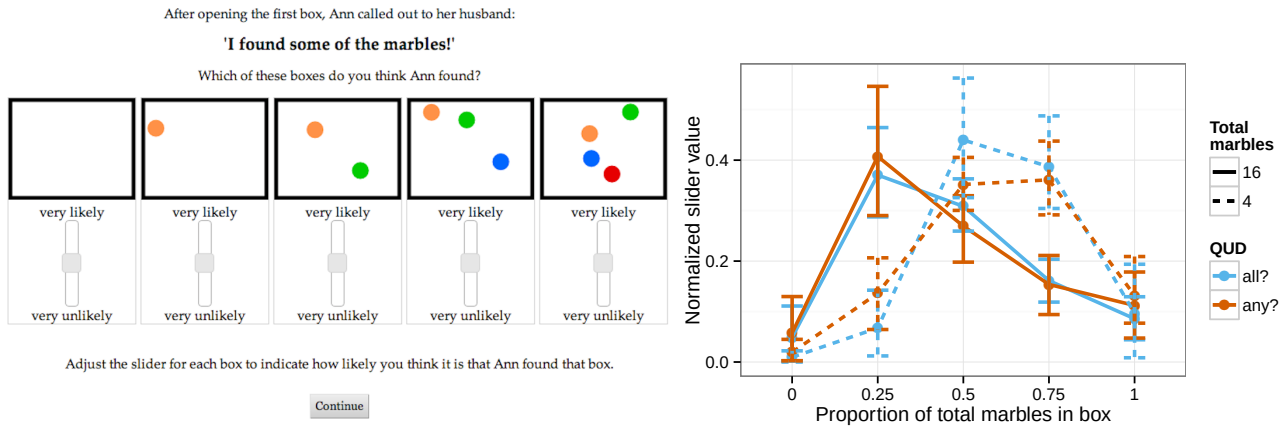

Figure 3: Left: Target display in Exp. 3, small set condition (4 total marbles). Right: Normalized slider values indicating probability of each box of marbles being the actual box in Exp. 3. Error bars indicate bootstrapped 95% confidence intervals.

or explicitly evaluating truth of utterance alternatives—do exhibit sensitivity to the contextual manipulations. Why is this? In the following we offer some speculative remarks about the nature of the different measures, their likely connection, and ideas for future work.

One possibility for why the contextual effects do not show up in the comprehension measure is that comprehension is simply more difficult or more noisy than production. This by itself is not a satisfying explanation without an argument for why comprehension should be more difficult—particularly in light of the ecological importance of comprehension and intrinsic relation between comprehension and production (it would not be communicatively useful for a speaker's production to be sensitive to context if a listener ignores context).

Another possibility is suggested by recent developments in computational modeling of pragmatic inference (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Franke, 2009; Franke & Jäger, 2013), which provides an intuitive but precise way to characterize the points in participants' reasoning that different measures probe. These models are based on the idea that listeners arrive at an interpretation by reasoning about what a speaker who is trying to be informative would have said. That is, listeners are modeled as having a model of the speaker's contextual utterance probabilities (in our case the probability of uttering *some*, *all*, etc.) that they invert using Bayes' rule (Equation 1 below) to recover the most likely states of the world (in our case, different possible boxes of marbles) from the speaker's actual utterance.

This perspective suggests that sentence interpretation tasks measure listeners' ultimate interpretation, while sentence verification and word probability rating measure listeners' underlying speaker models. If we take seriously the idea that the final listener interpretations are derived from Bayesian reasoning about the speaker model, the context effects observed at the speaker level may simply be too small to survive the final step of reasoning, leading to a null effect at the listener level. To validate this argument we next simulate the expected interpretation effect size (Exp. 3) from the production probabilities obtained in Exp. 2, the probabilistic model, and different assumptions about prior probabilities.

A simulation: the importance of the prior

For the sake of simplicity, we simulate only the effect of the QUD and ignore set size differences.[2] We follow previous probabilistic models of pragmatic inference (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) in modeling listener behavior by assuming that listeners can use Bayesian inference to recover the state of the world b (a box) that the speaker intended to convey, given that he produced utterance w. We further extend previous models by assuming that listeners condition their inference on the QUD. By Bayes' rule:

$$P_{\text{listener}}(b|w,\text{QUD}) \propto P_{\text{speaker}}(w|b,\text{QUD})P(b) \quad (1)$$

where $P(b)$ captures the listener's prior beliefs about the box and $P_{\text{speaker}}(w|b,\text{QUD})$ describes the listener's model of the words a speaker will choose. We assume only two box types $b = \{b_\exists, b_\forall\}$, where $b_\exists$ is a box that contains at least one but not all of the marbles and $b_\forall$ is the box containing the complete set of marbles.[3] We further assume there are four possible utterances $w = \{w_{\text{some}}, w_{\text{all}}, w_{\text{none}}, w_{\text{four}}\}$, which are the utterance alternatives participants saw in Exp. 2.

We take the empirical normalized mean slider values from the Exp. 2 small set condition as $P_{\text{speaker}}(w|b_\forall,\text{QUD})$. Since we did not include an incomplete set target display in Exp. 2, we instead set values for $b_\exists$ for both QUDs at *some*: 0.94, *all*: 0.02, *none*: 0.02, *four*: 0.02. We vary the prior probability $p(b_\forall)$ from 0.1 to 0.9 in steps of 0.1 and generate the listener posterior $P_{\text{listener}}(b|w,\text{QUD})$.

[2] We focus only on the small set condition here but the results are qualitatively the same for the big set condition.

[3] The set sizes 1 - 3, which form an equivalence class, are included in $b_\exists$. Alternately, one could model each set size independently.

Results are shown in Figure 4.[4] For the two different QUDs, the difference in the posterior probability of believing that the speaker intended to communicate the box containing the complete set is negligible for very small prior probabilities of $b_\forall$. With low $p(b_\forall)$, the predicted $p(b_\forall|w_{\text{some}}, \text{QUD})$ is small for either QUD, resulting in a floor effect. However, the predicted QUD effect increases as $p(b_\forall)$ increases. This provides a potential explanation for the observed lack of QUD effect on interpretation (Exp. 3): the prior probability of the complete set may have been too small to detect an effect.

Given the empirical comprehension means and confidence intervals from Exp. 3, the range of predicted $p(b_\forall)$ can be read off directly from Figure 4: the predicted prior range is the area where the model prediction confidence intervals for each QUD intersect the confidence intervals from the empirical data. This yields a predicted $p(b_\forall)$ between 0.22 and 0.66.

This result makes two predictions: a) measuring $b_\forall$ explicitly should yield values in the predicted range;[5] and b) increasing $p(b_\forall)$ experimentally should lead to an increase in the size of the QUD effect. We leave this step for future work.

Note that in one of the few reported cases of the successful use of the interpretation measure to detect an effect of context (in this case, of speaker knowledge) on scalar implicature, the prior for the complete set was explicitly manipulated to be large (Goodman & Stuhlmüller, 2013). This is consistent with the simulation reported here that predicts that a large prior makes it more likely to detect a contextual effect.

Conclusion

We have shown that the choice of dependent measure in experimental pragmatics research can greatly affect the conclusions drawn. Focusing only on the sentence interpretation measure would have led to the conclusion that there is no evidence to support contextual QUD and subitizing effects on scalar implicature. However, the production-based dependent measures did reveal contextual effects. We then showed that this asymmetry is predicted by models of pragmatic inference that assume listeners reason about informative speakers.

This work has both methodological and theoretical implications. Methodologically, that different ways of probing the underlying process are sensitive to these effects in different ways suggests that the choice of dependent measure is a serious issue that researchers doing experimental pragmatic research should take into account. Theoretically, this paper makes two contributions. First, it adds to a growing body of work showing that the strength of scalar inferences is modulated by an implicit Question Under Discussion and the estimated speaker effort of reducing uncertainty about set size.

[4] Confidence intervals on model predictions were obtained by fitting the model to the maximum and minimum of the bootstrapped 95% confidence intervals obtained from the empirical data.

[5] A preliminary pilot study, a minor variant of Exp. 3 in which participants were not shown the speaker's utterance, revealed that participants' prior $p(b_\forall)$ was between .04 and .24, consistent with the predicted value. For the time being, the empirical prior and the robustness of the predicted prior to varying model assumptions await a more systematic investigation.

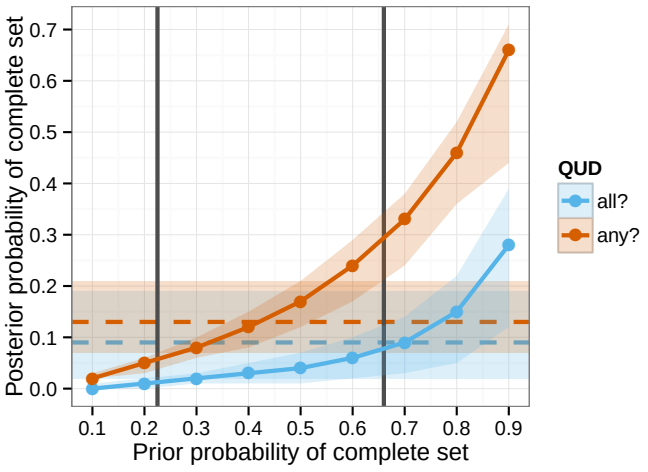

Figure 4: Model predictions for $p(b_\forall|w_{\text{some}}, \text{QUD})$ as a function of $p(b_\forall)$. Dashed lines/shaded areas indicate empirical means/confidence intervals from Exp. 3. Vertical lines indicate range of $p(b_\forall)$ consistent with observed data.

Second, it provides evidence for the utility of using formal, probabilistic models of pragmatics to understand the effects of choosing a dependent measure to probe effects of context on pragmatic inference.

References

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

Degen, J. (2013). *Alternatives in Pragmatic Reasoning*. Unpublished doctoral dissertation, University of Rochester.

Degen, J., & Tanenhaus, M. K. (in press). Processing scalar implicature: A Constraint-Based approach. *Cognitive Science*.

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

Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. Unpublished doctoral dissertation, Universiteit van Amsterdam.

Franke, M., & Jäger, G. (2013). *Pragmatic Back-and-Forth Reasoning*. Manuscript, Amsterdam-Tübingen.

Geurts, B. (2010). Quantity implicatures.

Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics and Pragmatics*, *2*, 1–34.

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

Roberts, C. (2004). Information structure in discourse. *Semantics and Pragmatics*, *5*, 1 – 69.

van Tiel, B. (2013). Embedded Scalars and Typicality. *Journal of Semantics*, 1–31.

Zondervan, A. (2010). *Scalar implicatures or focus: an experimental approach*. Unpublished doctoral dissertation, Universiteit Utrecht, Amsterdam.