Calculating scalar inference under QUDs*

Eszter Ronai & Ming Xiang

The University of Chicago

1. Introduction

Implicatures serve as an important testing ground for examining the processing cost of integrating semantic and pragmatic information. Starting with Bott and Noveck (2004), several studies have found that implicature computation is costly, evidenced by e.g. longer reaction times. More recently, attention has shifted towards identifying contextual cues and constraints that modulate this processing cost (Degen and Tanenhaus 2015). The exact source of the processing cost has also been investigated, with van Tiel and Schaeken (2017) proposing that it depends on the structural characteristics of the required alternatives. Specifically, retrieving (scalar) alternatives from the lexicon is argued to be what makes implicature calculation costly. In this paper we look at the calculation and processing of *scalar inference*, but crucially also manipulate context. Experiment 1 presents an elicitation study establishing what the relevant Questions Under Discussion (QUDs) are. Experiment 2, a sentence-picture verification study, shows that QUDs modulate calculation rates and processing speed. We find that under QUDs that bias towards calculating scalar inference, there is no increase in reaction times, but under QUDs that bias against calculating the inference we observe longer reaction times. Our results are thus most compatible with a constraint-based account of implicature, where QUD is one of many cues. Additionally, they provide evidence against an alternative-based account of processing cost.

2. Background

In natural language communication, hearers regularly infer messages beyond what is literally, explicitly said by the speaker, and in doing so, they rely not only on what is said, but also on what is not said. One well-studied instantiation of this phenomenon is (the family) of implicatures, exemplified in (1) by scalar inference.

*We would like to thank Chris Kennedy for his support throughout the whole project, Bob van Tiel for the experimental materials, Zsolt Veraszto for technical help, and audiences at AMLaP24 and NELS49 for their interest and helpful discussion. All remaining errors are our own.

(1) Mary ate some of the cookies.

Literal: Mary ate some and possibly all of the cookies.
Inference-enriched: Mary ate some but **not all** of the cookies.

To derive the inference-enriched reading of the sentence in (1), hearers consider and reason about the informationally stronger alternative that could have been uttered in place of what has actually been uttered. In particular, hearers are taken to reason that the stronger statement (*Mary ate all of the cookies.*) was also available to the speaker, but because she chose not to utter it, its negation can be inferred. This process can be viewed as an interaction of the Quality and Quantity maxims (Grice 1967).

2.1 The cost of implicature processing

One of the major questions in psycholinguistic studies of semantics-pragmatics is how fast implicatures are processed: do they arise generally as a default, or does generating them incur a cost? Much existing work has indeed shown that the calculation of scalar inferences incurs a processing cost, as evidenced by increased reaction times (Bott and Noveck 2004), ERP data (Noveck and Posada 2003), or delays in eye-tracking (Huang and Snedeker 2009). These results support a Literal-first model of implicature processing (Huang and Snedeker 2009). At the same time, however, some studies found evidence for a speedy implicature calculation process (Grodner et al. 2010) – lending support to the Deafult hypothesis about implicature processing (Levinson 2000).

On the other hand, instead of making a categorical distinction between "costly" vs. "cost-free" processing, a constraint-based approach views implicature calculation and processing as resulting from the interaction of multiple cues and constraints. Cues that have been shown to modulate the rate or speed of inference calculation include e.g. the syntactic partitive, or the availability of lexical alternatives (Degen and Tanenhaus 2015); the relevance of the stronger alternative proposition (Breheny et al. 2006, Politzer-Ahles and Fiorentino 2013); cognitive load (De Neys and Schaeken 2007); the speaker's knowledge state (Goodman and Stuhlmüller 2013); or face threat (Bonnefon et al. 2009). As mentioned, such findings are consistent with a cue-based or probabilistic model (e.g. Degen and Tanenhaus 2015's Constraint-Based framework, or Frank and Goodman 2012's Rational Speech Act models), which privileges neither semantics, nor pragmatics in processing.

Van Tiel & Schaeken (2017) have probed the exact source of the cost associated with implicature processing. Building on Katzir (2007) and Chemla and Bott (2014), they posit the Lexical Access hypothesis: implicature processing incurs a cost due to the retrieval of items from the lexicon in order to construct the alternative(s) relevant in the calculation of the implicature. Thus, under Lexical Access, what leads to a delay is not implicature processing or the counterfactual reasoning involved therein per se. Processing costs occur only if the lexicon needs to be accessed to construct alternatives. For example, in scalar inference (1), *all* is retrieved to produce the alternative *Mary ate all of the cookies*, ultimately leading to the computation of the inference *Mary ate some but not all of the cookies*. Van Tiel and Schaeken's (2017) experimental findings were in line with the Lexical Access hypothesis:

of four types of conversational implicatures tested, it is only scalar inference (*some-all*) that resulted in processing cost, the other three (which do not involve the retrieval of lexical alternatives) did not. We note, however, that van Tiel and Schaeken's (2017) study presented target sentences in the absence of any context. Much previous research has shown that context has a strong effect on the rate of implicature calculation, raising the question of whether it would also have an impact on processing, and if so, how that interacts with the cost of alternative construction. We thus turn to the role of context now.

2.2 The relevance of context

It has long been noted that depending on context, otherwise predicted implicatures can fail to arise. One way of theorizing about context has been with reference to the concept of QUDs, defined as the immediate topic of discussion, proferring a set of relevant alternatives (Roberts 1996/2012). Discourse is construed as giving rise to a stack of QUDs, and the ultimate discourse purpose is to answer all of these QUDs. An assertion is felicitous, then, if it chooses among the proferred alternatives and thereby bears upon the QUD. The below example from Levinson (2000) exemplifies the QUD-dependence of scalar inference:

(2) A: Is there any evidence against them?
B: Some of their identity documents are forgeries.
Predicted implicature: Not all of their identity documents are forgeries.

The predicted implicature in (2) would be consistent with the common ground and B's utterance, but it does not arise because A's question suggests that she is only interested in whether there is at least some evidence against the criminals. Thus there is no reason to consider *All of their identity documents are forgeries* as an alternative B could have said.

QUDs have also been experimentally shown to induce variation in scalar inference calculation rates. Experiments have been carried out either using an explicit QUD (Zondervan et al. 2008), or promoting one implicitly via a background story (Degen 2013) or via focus intonation (Cummins and Rohde 2015). Zondervan et al. (2008) and Degen (2013) investigated *some but not all* scalar inferences, placing them under QUDs containing *some* vs. *all* and *none* vs. *all* respectively. They found that reliably more inferences were calculated when an *all*-QUD is promoted. Cummins and Rohde (2015) found comparable results testing a wider variety of scalar inferences, e.g. *warm-not hot*, which is especially important in light of the recent finding that there is substantial variability among scales such as *some-all* vs. *warm-hot* with regard to the availability of the corresponding implicature (i.e. scalar diversity, van Tiel et al. 2016). Such findings constitute empirical confirmation that hearers do not always calculate scalar inference; rather they are more or less likely to do so depending on context. What has proven a challenge, however, is tracking down the QUDs relevant for a given context. Previous research has largely relied on the experimenters' own intuitions about what constitutes a likely QUD to an utterance, guided by linguistic theory, e.g. what the members of a lexical scale are (*none-some-all*).

In this paper we go beyond previous studies in empirically eliciting QUDs (Experiment 1), which constitutes a first attempt to address the problem of systematically identi-

fying QUDs relevant to a given context. In Experiment 2, we hypothesize that it is not only calculation rates, but also processing cost that tracks the QUD. Specifically, we attribute particular importance to supportive vs. non-supportive contexts, and predict that the likelihood of scalar inference calculation, and also whether there is an increase in reaction times, is tied to the context that the sentence occurs in. Our findings confirm our predictions.

3. Experiment 1: QUD elicitation

In this section we present an elicitation experiment, which established the likely QUDs for a given context for scalar inference (SI).

3.1 Participants

40 monolingual speakers of American English, recruited on Amazon Mechanical Turk, participated in an online elicitation experiment, administered on the Ibex platform (Drummond 2007). Participants were compensated $2.00.

3.2 Task, materials and procedure

The task was a modified sentence-picture verification task used for elicitation. Participants were provided with a background story that two people, Anne and Bob, are discussing pictures about shapes. Anne cannot see these pictures, so she is always asking Bob about what he sees. The instructions also emphasized that Bob always answers truthfully. (3) shows the instructions given to participants before the start of the experiment.

(3) In this experiment you are going to see dialogues between Anne and Bob, who are discussing pictures. Each picture shows a number of colored shapes.
However, only Bob can see the pictures, Ann cannot. So Anne is always asking Bob about what he sees.

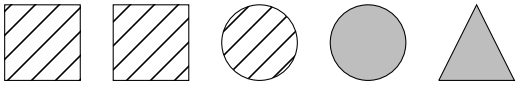
Participants then saw written SI target sentences paired with pictures, and were told that the sentences were Bob's answers to Anne's questions. Target sentences and pictures were adapted from the materials used by van Tiel and Schaeken (2017), in order to make our results directly comparable to theirs, with the addition of the context manipulation. The target sentences investigated were of the following form: *Some of the shapes are blue* (intended inference: Not all of the shapes are blue). In place of Anne's question participants saw a blank line - see (4) for an example of a trial screen. Participants were instructed to guess what Anne's question could have been, based on Bob's answer and the picture, and had to provide their response in writing.

(4) Anne: _____*?*
Bob: *Some of the shapes are blue.*
Picture: Good Control or Target

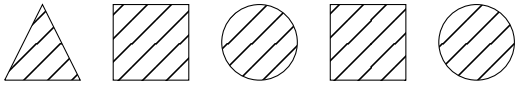
There were two types of pictures: Bob's answers were unambiguously good descriptions of Good Control pictures. For Target pictures, they were good descriptions on their literal (*Some and possibly all of the shapes are blue*), but not on their inference-enriched reading (*Some but not all of the shapes are blue*). Examples of the pictures are in (5).

(5) *Picture stimuli in Experiment 1*[1]

Control: Good



Target: Underinformative



Good Control pictures always contained five shapes, three of which matched the color mentioned in Bob's sentence (here: blue/striped), while two were a different color (here: yellow/shaded). Target pictures contained five shapes of the same color, the one mentioned in the sentence. Shapes were varied between triangle, circle and square (with pictures containing either a mix of shapes or the same shape five times), and colors were varied between blue, yellow, red, green, orange and black.

The experimental design included the two-level factor of Picture type as a between-participants manipulation: 20 participants saw only Good Control, and 20 other participants only Target pictures. Each participant saw 15 SI trials.

3.3 Results and discussion

Though thirteen distinct kinds of questions were elicited, the majority of answers came from a much smaller set. (6) provides examples of the most commonly offered QUD types and their respective frequencies in the data.

(6) Dominant question types and their frequencies with the two pictures:

Question type	Target	Good Control
What color are the shapes?	42%	32%
Are **any** (of the) shapes blue? Are there (**any**) blue shapes?	25%	33%
Are **all** of the shapes blue?	6%	20%
Are **some** of the shapes blue?	12%	2%

We can see that a small number of questions dominated in the elicitation task, with *any*-QUDs and *what*-QUDs being roughly equally frequent. Another observation to be made about the data is that for both Good Control and Target pictures, the same kind of questions have been elicited and in largely the same frequencies. This suggests that both pictures, and therefore both the literal and inference-enriched interpretations of the sentences, are

[1] In this paper we illustrate color with patterns, such that e.g. "blue" is represented by "striped".

independently compatible with the elicited questions. This is somewhat puzzling, considering that several of the elicited questions were previously thought to bias towards one or the other interpretation. For example, *all*-QUDs have been argued (and experimentally shown) to bias towards the enriched, and *any*-QUDs towards the literal reading. Based on this, we might predict that only Good Control pictures, which are compatible with the inference-enriched meaning of SI, would have elicited e.g. *all*-questions. Conversely, we might not have predicted Target pictures, which are not compatible with the enriched reading, to elicit *all*-questions. Yet we find that biasing questions are still generally compatible with both types of pictures. One thing to keep in mind is that in this experiment, we do not obtain information about what interpretation participants actually assigned to the target sentences. As many studies (including our Experiment 2) have found, some participants do judge *Some of the shapes are blue* to be a good description of both Good Control and Target pictures – in these cases, it is perhaps unsurprising that both pictures elicited the same type of questions. Future research should further probe these issues about the interplay of production and interpretation.

The primary aim of Experiment 1 was to test in a more systematic and empirically grounded manner what the likely QUDs are for scalar inference. Previous work on the role of context in SI calculation has largely assumed questions that contain members of the relevant lexical scale. We can see that our elicitation study has indeed uncovered questions that have not been discussed in existing literature as being relevant to implicature derivation or the particular target sentences, e.g. the *what*-QUD. On the other hand, some questions previously discussed in theoretical or experimental contexts did in fact show up in our elicitation data, e.g. *any*-, *some*- and *all*-QUDs. We note, however, that they appeared less frequently than one might predict, and as mentioned, we also did not find a robust bias towards either the Target picture (promoting the literal reading), or the Good Control picture (promoting the inference-enriched reading) with any of the questions.

Nonetheless, Experiment 1 established likely QUDs whose effects on SI computation we can then investigate. We therefore took the most frequent questions from Experiment 1, and used them in Experiment 2 to see if they modulate implicature calculation rates and processing cost.

4. Experiment 2: QUD manipulation

We present a sentence-picture verification experiment, where we embedded scalar inference sentences under the most frequent QUDs elicited in Experiment 1. Our results showed that both the probability of inference calculation and also whether there is a delay in reaction times are conditioned on the QUD.

4.1 Participants

90 (30 in each QUD condition) native monolingual speakers of American English, recruited on Amazon Mechanical Turk, participated in the experiment, administered on the Ibex platform (Drummond 2007). 5 participants were removed from the analysis, due to having a mistake rate exceeding 25% on filler items. Participants were compensated $1.50.

4.2 Task, materials and procedure

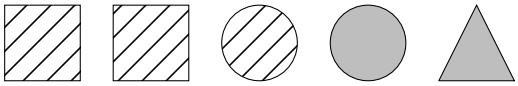
Experiment 2 employed the sentence-picture verification paradigm, with target sentences embedded in a dialogue context. Participants were given the same background story as in Experiment 1: Anne is asking questions from Bob, about pictures that only Bob can see. On the first screen, participants saw Anne's question and were instructed to press a key after reading it. On the following screen, they saw Bob's answer to Anne's question, as well as the picture they were discussing. Participants were instructed to make a binary judgment (by clicking on a button) about whether Bob gave a good answer to Anne's question, given the picture he saw. The two buttons said "Good" and "Not Good". Participants' choices were recorded, as well as reaction times from the onset of the second screen (displaying Bob's utterance and the picture) until the participant pressed one of the buttons.

The experiment featured a three-by-three design: crossing Picture (within-participants: Good Control, Bad Control, Target) and QUD (between-participants: what, any, all). Similarly to Experiment 1, there were two types of pictures: Control, of which Bob's sentence was an unambiguously good/bad description, and Target, where the judgment depends on whether the inference has been derived - see (7). Bob's answers are good descriptions of the Target pictures on their literal, but not on their inference-enriched reading. Good Control and Target pictures were identical to those used in the elicitation experiment. Bad Controls were also added in Experiment 2. They contained five shapes, none of which has the color (here: blue/striped) mentioned in the sentence.

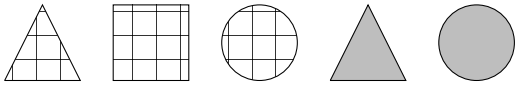
(7) *Picture stimuli in Experiment 2*

Bob: *Some of the shapes are blue.*

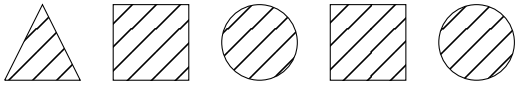
Control: Good



Control: Bad



Target: Underinformative



The QUD condition (i.e. Anne's questions) was also a three-level manipulation: *what*-, *any*- and *all*-QUDs. These questions were the most frequent elicited questions in Experiment 1 (see (6)). (8) provides an example of the question manipulation.

(8) QUD manipulation in Experiment 2

what: **What** color are the shapes?
any: Are there **any** blue shapes?/Are **any** shapes blue?
all: Are **all** shapes blue?

QUD was manipulated between participants, such that each participant was placed in either the *what*-, or the *any*-, or the *all*-QUD condition. To prevent fatigue effects due to only encountering the same type of question, experimental items were intermixed with filler items, which included different shapes and questions unrelated to SI and the question manipulation. Each participant saw 12 SI items.

4.3 Predictions

Based both on previous empirical results (Zondervan et al. 2008, Degen 2013) and theoretical proposals (Question-Answer Requirement, Hulsey et al. 2004), we can make the predictions that *all*-QUDs would result in the highest rate of implicature calculation. *Any*-QUDs, on the other hand (see e.g. (2)), are predicted to bias against deriving the implicature and lead to lower calculation rates. The prediction for *what*-QUDs is less clear, in part because they have previously not been treated as relevant QUDs in these contexts, and it is less obvious what existing theoretical frameworks would predict about them. Interestingly, however, they are the most frequent QUDs in the elicitation data (Experiment 1), and we treat their biasing behavior as an empirical question.

Implicature calculation rates are indexed by the proportion of "Good" responses to Target pictures: if a participant says that Bob gave a "Good" description of a Target picture, she has not calculated the SI inference. We thus predict variation across QUDs in the percentage of "Good" responses to Target. For example, *any*-QUDs are predicted to result in lower rates of calculation and therefore higher "Good" percentages for Target pictures. In what follows, we make a distinction between Literal-biasing QUDs, which lead to lower rates of inference generation (the prediction for the *any*-QUD) and Inference-biasing QUDs, which lead to higher rates of inference generation (the prediction for the *all*-QUD).

Crucially, we also predict that QUDs affect not only implicature calculation, but also processing, as operationalized by reaction times (henceforth RT). In line with previous literature, we take longer RT when responding "Not Good" to Target, relative to the RT when responding "Not Good" to (Bad) Control, to be what indexes the cost of implicature calculation. This is because Bad Controls can be rejected based on literal, semantic meaning, but the rejection of a Target picture suggests that the participant has gone through the inference calculation process. Our predictions, then, are that Inference-biasing QUDs (i.e. those that bias towards deriving the scalar inference) will not make implicature derivation incur a cost. That is, we should not see a difference in "Not Good" to Target as compared to "Not Good" to Control RTs in the Inference-biasing QUD condition(s). On the other hand, Literal-biasing QUDs (i.e. those that bias against deriving an implicature) will have the effect that when the implicature is derived, its calculation is a costly and therefore slower process. Under Literal-biasing QUDs, then, we predict an increase in the time it takes to respond "Not Good" to Target as compared to "Not Good" to Control.

4.4 Results and analysis

Figure (9) plots the proportions of "Good" responses for SI. The primary purpose of Good and Bad Control pictures was to make sure participants are adequately doing the task,

which we can see from responses being largely at ceiling and floor respectively. We thus focus on the analysis of the more informative Target pictures, which constituted our critical manipulation. Recall that for Target pictures, a "Good" answer indicates implicature non-calculation. For the statistical analysis, a logistic regression model (glm function in R) was fit (mixed effects models did not converge), predicting Response (Good vs. Not Good) by QUD. Because our main prediction concerns the *any*-QUD vs. *all*-QUD difference, levels within the QUD variable were treatment coded, with *any* serving as the reference level. The statistical analysis revealed that the difference between *all* vs. *any* ($p<0.001$) is significant, while *what* resulted in responses somewhere in between the two other QUDs, with the *any* vs. *what* ($p<0.06$) difference being only marginally significant.

Our prediction was that QUDs would modulate the rate of pragmatic inferences derived, as evidenced by rates of judging Bob's answer (as a description of the Target picture) "Good" vs. "Not Good". Our findings are in line with these predictions. We found significantly fewer implicatures calculated under *any*-QUDs than under *all*-QUDs: that is, *any* is a Literal-biasing, while *all* is an Inference-biasing QUD. The *what*-QUD falls in the middle, making it unclear whether it can be categorized as either Literal- or Inference-biasing.

(9) *Proportion of participants' "Good" responses by Picture condition. Different colors denote the different QUD conditions. Error bars represent standard error.*

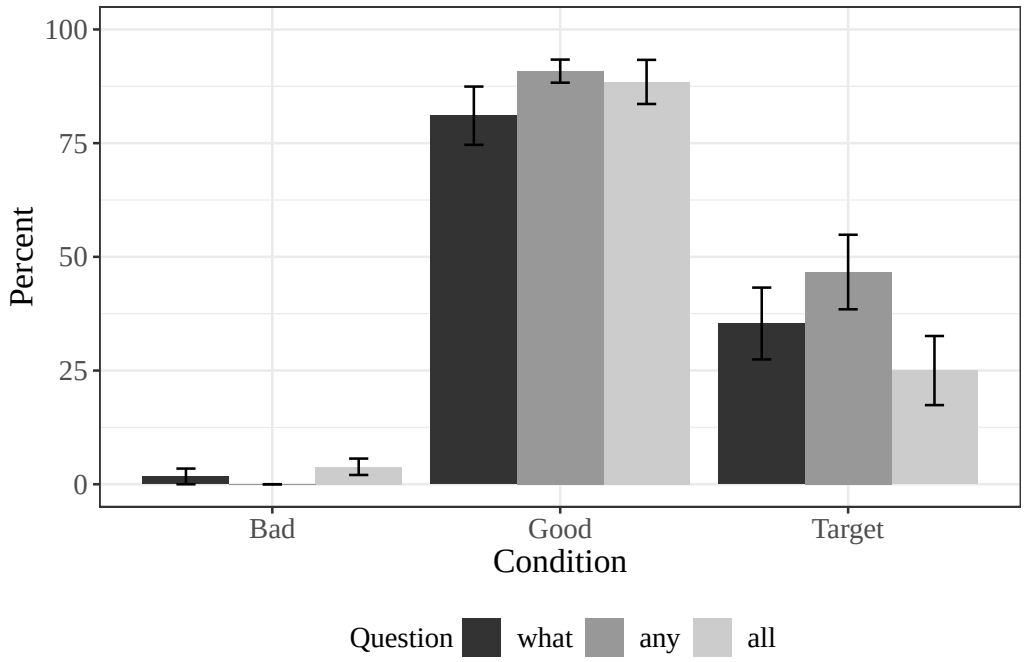
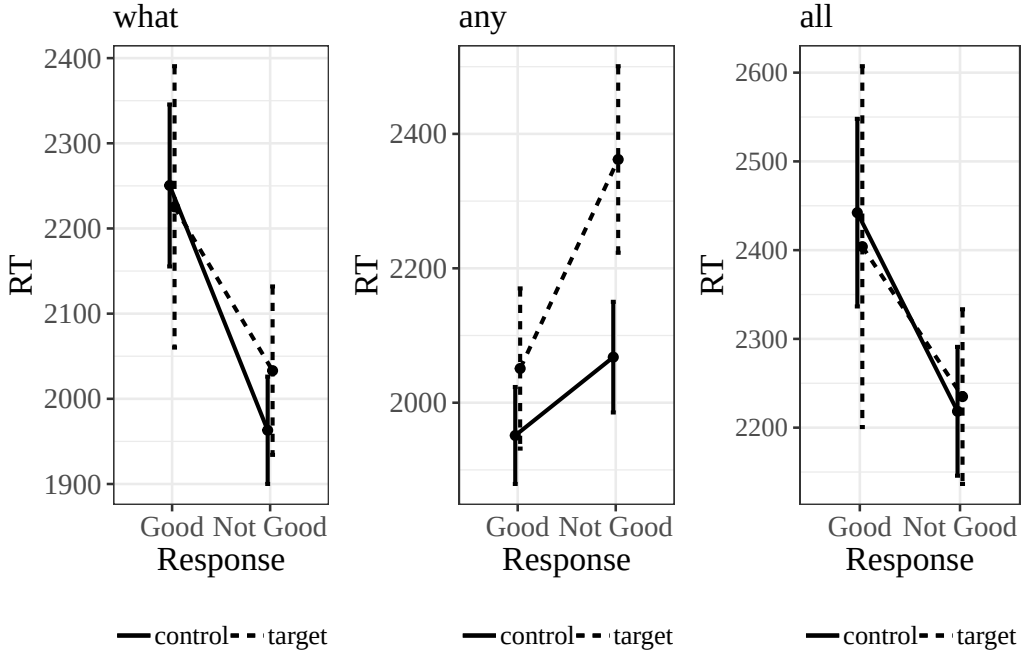


Figure (10) plots reaction times broken down by QUD. Because our predictions concern the difference in RT when responding "Not Good" to Target, as compared to Control (see Section 4.3), and we do not have specific hypotheses about RTs when responding "Good", we restrict the statistical analysis to "Not Good" responses. Nevertheless, the full data set is plotted in Figure (10). For the statistical analysis, a linear mixed effects model (lmer

from the lme4 package in R, Bates et al. 2015) was fit, predicting RT by Condition (Target vs. Control). Levels within Condition were treatment coded, with Control serving as the reference level. Random slopes and intercepts were included for participants and items. Whenever the full model did not converge, the random effects structure was simplified following the recommendations of Barr et al. (2013). In the following we analyze RT data question-by-question, following the predictions we made. The p values reported below were obtained using the lmerTest package in R (Kuznetsova et al. 2017).

(10) *Mean reaction times (RT) in ms for judging Bob's answer as "Good" or "Not Good". Different QUD conditions are on separate plots. Line type denotes Control (Good and Bad) vs. Target pictures. Error bars represent standard error.*



For the *any*-QUD, we found a significant difference in RT between responding "Not Good" to Target vs. Control ($\beta = 391.47$, $t = 2.599$, $p < 0.05$). In particular, RTs were increased for the Target picture, as compared to the Control picture when responding "Not Good" (see middle panel in Figure (10)). For the *all*-QUD, there was no significant difference in RT between responding "Not Good" to Target vs. Control ($\beta = 5.062$, $t = 0.043$, $p = 0.966$). As can be seen in the rightmost panel in Figure (10), the *all*-QUD did not lead to increased RTs for "Not Good" to Target as compared to Control. Similarly, for the *what*-QUD, we found no significant difference in RT between responding "Not Good" to Target vs. Control ($\beta = 194.91$, $t = 1.737$, $p = 0.0992$). That is, the *what*-QUD patterns with the *all*-QUD. This suggests that despite the mixed results in terms of SI calculation rate, the *what*-QUD is Inference-biasing based on reaction time measures.

In sum, our predictions are largely borne out in the data: the questions which we could unambiguously classify as Literal- (*any*-QUD), vs. Inference-biasing (*all*-QUD) show di-

verging behavior in terms of reaction time cost. *Any*-QUDs resulted in a reaction time cost, while *all*-QUDS (as well as *what*-QUDs) did not incur a cost for inference calculation. We discuss the implications of these results in the next section.

5. General discussion

Our hypothesis was that the QUD manipulation we introduced in Experiment 2 would reveal that some questions bias towards deriving the pragmatic inferences from the target sentences, while others bias towards their literal meaning. Indeed, we found variation across QUDs in the percentage of "Good" responses to Target. In particular, our choice proportion findings show that the *any*-QUD is Literal-biasing, i.e. it led to lower rates of implicature calculation. On the other hand, the *all*-QUD was found to be Implicature-biasing, i.e. it led to higher calculation rates. The *what*-QUD fell somewhere in the middle. Differences among QUDs also predicted reaction times. In particular, we found that with Literal-biasing *any*-QUDs, making an SI-enriched judgment took longer than responding to the relevant literal control. With the Inference-biasing *all*-QUD, however, there was no difference between the "Not Good" RTs, i.e. there was no reaction time cost for inference calculation. *What*-QUDs, which showed no clear pattern in the choice proportion data, turned out to be Inference-biasing in the RT data, and patterned with the *all*-QUD. These findings strongly suggest that SI computation is only costly when preceded by non-supportive QUDs.

The reaction time findings provide a challenge to a model of implicature processing that posits that literal meanings are always accessed first, and enriched meanings therefore come at a reaction time cost (Huang and Snedeker 2009). Similarly, they are also hard to accommodate under a model that assumes implicatures to always be calculated by default (Levinson 2000). Instead, they favor a model of implicature processing where neither literal, semantic meanings, nor inference-enriched, pragmatic meanings are privileged across the board (Degen and Tanenhaus 2015, Frank and Goodman 2012). Instead, the calculation and processing of implicatures arises from the interaction of contextual cues and constraints, of which QUD is one. Additionally, our findings are also informative with respect to the recent Lexical Access proposal of van Tiel and Schaeken (2017), which takes processing cost of implicatures to be triggered by lexical retrieval. Lexical Access would always predict a cost for SI calculation, because *all* is always retrieved to construct the relevant alternative. QUDs could introduce further modulation, but even under congruent QUDs, there should always be some residual cost triggered by lexical retrieval. This, however, is not what we find: under two of the three QUDs investigated, SI calculation did not lead to a reaction time cost. We therefore argue that, though open question remain, a QUD-based account better explains the current findings than an alternative-based account.

As for the specific mechanism of deriving our results under a QUD-based model, we can turn to e.g. the Question-Answer Requirement (QAR, Hulsey et al. 2004) approach:

(11) The Question Answer Requirement (QAR)
The selected interpretation of an ambiguous sentence, whether true of false, is required to be a good answer to the Question Under Discussion. (A good answer is an interpretation that at least entails an answer to the QUD.)

That is, any sentence is to be understood as an answer to a question, this question being the QUD (Hulsey et al. 2004, Gualmini et al. 2008). Our results can be accommodated under the QAR as follows. Using a standard Hamblin (1976) semantics, the meaning of a question, including the meaning of a QUD, is a partition of the set of possible worlds. Given that *any* is synonymous with *some*, the QUD *Are there any blue shapes?* partitions logical space (conceived of as the set of possible worlds) into two cells: one containing all the worlds where (12a) is true, and one containing all the worlds where (12b) is true.

(12) a. Some and possibly all shapes are blue.
b. No shapes are blue.

Given this QUD, then, deriving and not deriving the implicature both result in meanings that obey the QAR. It is not necessary to derive the SI inference from *Some of the shapes are blue*, because the literal interpretation *Some and possibly all of the shapes are blue* corresponds to (12a). Deriving the inference also results in a good answer in the sense of the QAR (i.e., an interpretation that at least entails an answer to the QUD), because the inference-enriched interpretation (*Some but not all of the shapes are blue*) of the sentence entails (12a). Therefore, under the *any*-QUD (i.e. Anne's question), Bob gave a good response whether or not the inference is derived.

The picture changes for the SI inference *some but not all* under the QUD *Are all shapes blue?*, which results in the following partitioning (assuming a five-shape display):

(13) a. All shapes are blue.
b. Not all shapes are blue.

In this case, only if the inference is derived is the dialogue QAR-compliant, because the literal reading does not address the QUD. However, the enriched reading *Some but not all of the shapes are blue* entails (13b). Thus, Bob only gave a good response to Anne's *all*-question on the inference-enriched reading, which is reflected in our finding that significantly more implicatures were derived with *all*- than with *any*-QUDs. The remaining question is why *what*-QUDs were very frequent in elicitation, although they are not straightforwardly accommodated under QAR – a discrepancy that future work should address.

6. Conclusion

The processing cost (measured via e.g. reaction times) of calculating implicatures such as scalar inference has been of great interest in psycholinguistic studies of semantics-pragmatics. Prior work has also shown that context (e.g. QUDs) modulates how likely pragmatic inferences are to be derived. In this paper we investigated the effect of QUDs not only on the calculation rates, but also the processing of *some-all* scalar inference. In Experiment 1, we took a first step in addressing the problem of narrowing down potential QUDs for a given context and conducted an elicitation study, the results of which fed into our QUD manipulation experiment. In doing so, we went beyond previous work that relied only on theory-informed intuitions about what may serve as a relevant QUD, and instead

we treated this issue as an empirical question. Novel data from Experiment 2 (QUD manipulation) showed that not only do QUDs modulate implicature calculation rates, they also affect the reaction time cost of processing. This challenges previous discussions of there being a uniform cost (or lack of cost) for deriving scalar or quantity implicatures. Instead, our findings are most compatible with a constraint-based account of implicature, and language processing more generally, where QUD is one of many cues.

Additionally, our findings address not only whether there is uniform cost or lack of cost in implicature processing, but also what the source of that cost is. In a recent proposal, van Tiel and Schaeken (2017) argued that this processing cost is tied to the characteristics and construction of the alternatives required in the reasoning process. Specifically, under Lexical Access, processing cost is directly triggered by alternative construction, predicting scalar inference calculation to always be costly. This is because in order to construct the relevant alternative (*all of the shapes* for *some of the shapes*), the lexicon needs to be accessed. This, however, is not what we found: only under the Literal-biasing *any*-QUD was a reaction time cost incurred. These findings strongly suggest that the processing cost of generating pragmatic inferences cannot be fully explained by alternative construction. The novel processing results showed that non-supportive questions result in a reaction time cost, but supportive questions do not – crucially even when lexical access occurs. Altogether, our results favor an account where processing tracks the QUD, instead of implicature calculation always incurring a cost/being cost-free, or cost being tied to alternative construction.

References

Barr, Dale J, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68:255–278.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1–48.

Bonnefon, Jean-François, Aidan Feeney, and Gaëlle Villejoubert. 2009. When some is actually all: Scalar inferences in face-threatening contexts. *Cognition* 112:249–258.

Bott, Lewis, and Ira A. Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language* 51:437–457.

Breheny, Richard, Napoleon Katsos, and John Williams. 2006. Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition* 100:434–463.

Chemla, Emmanuel, and Lewis Bott. 2014. Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition* 130:380–396.

Cummins, Chris, and Hannah Rohde. 2015. Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology* 6:1779.

De Neys, Wim, and Walter Schaeken. 2007. When people are more logical under cognitive load – Dual task impact on scalar implicature. *Experimental Psychology* 54:128–133.

Degen, Judith. 2013. Alternatives in pragmatic reasoning. Doctoral dissertation, University of Rochester.

Degen, Judith, and Michael K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive Science* 39:667–710.

Drummond, Alex. 2007. Ibex Farm. `http://spellout.net/ibexfarm`.

Frank, Michael C., and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336:998–998.

Goodman, Noah D., and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5:173–184.

Grice, Herbert Paul. 1967. Logic and conversation. In *Studies in the way of words*, ed. Paul Grice, 41–58. Harvard University Press.

Grodner, Daniel J., Natalie M. Klein, Kathleen M. Carbary, and Michael K. Tanenhaus. 2010. "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition* 116:42–55.

Gualmini, Andrea, Sarah Hulsey, Valentine Hacquard, and Danny Fox. 2008. The Question-Answer Requirement for scope assignment. *Natural Language Semantics* 16:205–237.

Hamblin, Charles Leonard. 1976. Questions in Montague English. In *Montague grammar*, ed. Barbara H. Partee, 247–259. Academic Press.

Huang, Yi Ting, and Jesse Snedeker. 2009. Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology* 58:376–415.

Hulsey, Sarah, Valentine Hacquard, Danny Fox, and Andrea Gualmini. 2004. The Question-Answer Requirement and scope assignment. 71–90. MITWPL.

Katzir, Roni. 2007. Structurally-defined alternatives. *Linguistics and Philosophy* 30:669–690.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82:1–26.

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. The MIT Press.

Noveck, Ira A., and Andres Posada. 2003. Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language* 85:203–210.

Politzer-Ahles, Stephen, and Robert Fiorentino. 2013. The realization of scalar inferences: Context sensitivity without processing cost. *PLOS ONE* 8:1–6.

Roberts, Craige. 1996/2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5:1–69.

van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33:137–175.

van Tiel, Bob, and Walter Schaeken. 2017. Processing conversational implicatures: alternatives and counterfactual reasoning. *Cognitive Science* 41:1119–1154.

Zondervan, Arjen, Luisa Meroni, and Andrea Gualmini. 2008. Experiments on the role of the question under discussion for ambiguity resolution and implicature computation in adults. In *Proceedings of Semantics and Linguistic Theory (SALT) 18*, ed. Tova Friedman and Satoshi Ito, 765–777.

Eszter Ronai, Ming Xiang
ronai@uchicago.edu, mxiang@uchicago.edu