Why do you ask? To be informative.

Robert X. D. Hawkins (rxdh@stanford.edu)

Andreas Stuhlmüller (andreas@stuhlmueller.org)

Judith Degen (jdegen@stanford.edu)

Noah D. Goodman (ngoodman@stanford.edu)
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

Keywords: questions; answers; computational pragmatics; theory of mind;

Introduction

Suppose your friend approaches you and asks, "Who is coming to the concert tonight?" How do you respond? You certainly don't need to give the full list of attendees – most of whom you do not know – even though they would all technically be valid answers. Instead, you might only mention the set of *mutual acquaintances* you know are planning to come, assuming that your friend doesn't care about the rest of the crowd. Now, imagine a different scenario. Suppose that you're waiting in line at the box office and want to find out whether your acquaintances had tickets as well. What question would you ask to the person in the ticket booth? If you asked, "Who is coming to the concert tonight?" you would likely get a quizzical look and an answer like "I don't know, a lot of people, why?" Instead, you might have to directly ask about your friends.

Since both questioners and answerers are already acutely sensitive to one another's intentions and knowledge, what makes a question useful? What makes an answer to a question useful? In this paper, we present three progressively more sophisticated computational models of question-answer behavior, which formalize and probe this deep interaction between the way answerers infer intentions and the way questioners signal them. We compare these models on the basis of two simulations of classic question-answer phenomena and one experiment in which participants must ask and answer questions given a fixed set of goals. We find that a sophisticated pragmatic answerer is needed to account for the data, and close by proposing that the purpose of questions in dialogue is to provide cues to the answerer about the questioner's goals and intentions.

A number of studies in psycholinguistics have provided evidence that answerers are both sensitive to a questioner's goals and attempt to be informative with respect to those goals. For example, when people are asked 'Do you have the time?" they typically round their answers to the nearest 5 or 10 minute interval, even when they're wearing a digital watch (Der Henst, Carles, & Sperber, 2002). However, if the question is preceded by the statement "My watch stopped," people make their response precise to the minute (Gibbs Jr & Bryant, 2008).

Similar evidence comes from a classic study where researchers called liquor merchants and asked, "Does a fifth of Jim Beam cost more than $5?" If this was preceded by the statement, "I want to buy some bourbon," merchants gave the actual price significantly more frequently than when it was preceded by the statement, "I've got $5 to spend." In the former case, the merchant inferred that the questioner's goal was just to buy whiskey, so the exact price was the maximally relevant response. In the latter case, the merchant inferred that the questioner's goal was literally to find out whether or not they could afford the whiskey, hence a simple 'yes' sufficed (Clark, 1979). Context and questioner goals have also been implicated in accounts of identification questions like "who is X?" (Boër & Lycan, 1975), and to questions like "where are you?" that permit answers at many levels of abstraction (Potts, 2012).

Recent formal models of question-answer pragmatics have made progress by incorporating questioner goals, and specifying what it means for an answerer to be informative with respect to these goals. van Rooy (2003), for instance, formalizes a goal as a utility function defining a decision problem faced by the questioner. A useful answer under this decision theoretic account is one that maximizes the expected value of the questioner's utility by reducing their uncertainty about the true state of the world. A useful question is one that allows for a sufficiently fine-grained set of answers, optimally distinguishing the worlds relevant to their decision problem. While this framework elegantly accounts for the context-dependence and relevance-maximization of question and answer behavior, it requires that the questioner's decision problem is known *a priori* by the answerer or fully determined by context. This minimizes the role of the questioner; they just establish a space of answers and then let the answerer do the rest of the work. If the answerer is so adept at using context to determine the relevant information, though, why does the questioner need to ask a question in the first place? Is it just a formality, to prompt the other for their information, or does it serve as a signal in itself?

We claim that the questioner must reason about answerer behavior, in order to determine what question will produce the most useful answers. This raises a further issue: what

kind of answerer does the questioner reason about, and is this internal model accurate? The rest of this paper is structured as follows. First, we specify a family of questioner and answerer agents extending the Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013), highlighting some points of divergence from previous RSA models. We then individuate three particular models in this family, representing progressively more sophisticated hypotheses about how questioners and answerers reason about their task. In particular, we compare a pragmatic answerer making inferences about the questioner's goals to two simpler models: one that takes into account only that an answerer wants to be maximally informative with respect to the explicit question asked (without inferring the questioner's underlying decision problem) and one that provides a literal answer to the question (without attempting to be maximally informative).

To compare these models, we derive predictions for a pair of experiments using a novel guessing-game task, and compare these predictions to human performance. In one phase of the task, we require participants to ask a question (from a fixed set of possible questions), given a decision problem. In the second phase, we require participants to give an answer (from a fixed set of possible answers) to a question (from a fixed set of possible questions). These models and data, combined with the psycholinguistic data above, seeks to place question-answer behavior in the larger class of social behavior governed by theory of mind.

A Rational Speech Act model of question and answer behavior

Suppose there is a set of distinct world states $\mathcal{W}$, a set of possible goals $\mathcal{G}$, a set of possible questions $\mathcal{Q}$, and a set of possible answers $\mathcal{A}$. These sets are all taken to be in common ground. We begin by thinking about how an optimal questioner should choose a question. Critically, instead of trying to impart information about the state of the world, a questioner attempts to *learn information about a private goal*, sometimes called a QUD (or question under discussion) (Roberts, 1996). A goal $g \in \mathcal{G}$ is a projection function that maps a complete world state to a particular feature or set of features that the questioner cares about. Each of these projections corresponds to a different utility function, in a decision-theoretic formulation. In order to learn information about their private goal g, the questioner reasons about how an internal model of an answerer would respond given some true world.

- The **questioner** takes a goal $g \in \mathcal{G}$ as input and returns a distribution over questions $\mathcal{Q}$. To do this, it first computes a prior $P_g(w)$ over the features of the world relevant to its goal. For each question $q \in \mathcal{Q}$, it computes the expected information gain, averaged over all possible true worlds, and all possible responses the answerer could give. Information gain is measured as the Kullback-Leibler divergence between the prior distribution and the posterior distribution over world states after hearing the answerer's (hypothetical) response:

$$P(q|g) \propto \sum_{a \in \mathcal{A}} \sum_{w^* \in \mathcal{W}} P(w^*) P(a|q,w^*) D_{KL}(P_g(w) \,\|\, P_g(w|q,a))$$

This questioner depends critically on the answer distribution $P(a|q,w^*)$, which it uses to upweight questions q that elicit useful answers across different world w^* and downweight questions that elicit vague or irrelevant answers. It also depends on $P_g(w|q,a)$, an 'interpreter' function that gives the likelihood of different worlds given question and answer pairs. We will specify this function later. First, we propose three different answerer agents that embody different assumptions that the questioner could make.

- The **literal answerer** takes a question utterance $q \in \mathcal{Q}$ and a true world state $w \in \mathcal{W}$ as input and returns a distribution over the answer space $\mathcal{A}$. It samples an answer a from $\mathcal{A}$ with prior probability $P(a)$ and conditions on the likelihood of the questioner inferring the true world w from this answer, using only the interpreter function as a cue. Note that this agent ignores the question; it's only concern is to convey the true state of the world. This makes it equivalent to the speaker in previous RSA models.

$$P(a|q,w^*) \propto P(w = w^*|q,a)P(a)$$

- The **explicit answerer** samples an answer a from $\mathcal{A}$ with prior probability $P(a)$, then uses the explicit utterance q as a QUD when evaluating the likelihood of a world under an interpreter.

$$P(a|q,w^*) \propto P_q(w = w^*|q,a)P(a)$$

- The **pragmatic answerer** uses an internal model of the explicit questioner (i.e. the questioner who reasons about an explicit answerer) as a generative model of questions given goals in order to estimate the likelihood of different goals given the question q. It samples an answer a from $\mathcal{A}$ with prior probability $P(a)$, uses the internal questioner model to estimate the likelihood of different goals $g \in \mathcal{G}$ given their question q, and attempts to be informative with respect to this goal:

$$\begin{aligned} P(a|q,w^*) &\propto P(g|q) \times P_g(w = w^*|q,a) \times P(a) \\ &\propto P(q|g)P(g) \times P_g(w = w^*|q,a) \times P(a) \end{aligned}$$

Finally, we must define the interpreter function that all of these agents are using to compute the likelihood of a world given a question and an answer. This is where we specify semantic assumptions about the meaning of a question and answer. For the purposes of this paper, we will use Groenendijk & Stokhof semantics (1984), where a question induces a partition $\mathcal{P}_q$ over the space of possible world and each cellof this partition is an equivalence class corresponding to a different answer. An answer, then, selects a cell of this partition, denoted by $\mathcal{P}_q(a)$, which is a set.

- The **interpreter** takes a question q and an answer a and returns a distribution of worlds that are consistent with this pair:

$$P(w|q,a) = P(w)\mathbb{I}_{\mathcal{P}_q(a)}(w)$$

where $\mathbb{I}_A(w)$ is the indicator function returning 1 if $w \in A$ and 0 otherwise.

This concludes our specification of the model space, giving a set of three answerers and three corresponding questioners that reason about them. Within this computational framework, different theories of question and answer behavior can be formalized and compared on the basis of their predictions. Assumptions about what is held in common ground are made transparent, and we can systematically manipulate individual elements of the model to test how they affect overall predictions. Because these are probabilistic models, we can succinctly write them down and evaluate their predictions in a probabilistic programming language (Goodman, 2013). The model predictions shown throughout the rest of the paper were computed using a language called WebPPL (Goodman & Stuhlmüller, electronic). Note that this specification diverges from previous work in the RSA framework in a few key ways: (1) for the questioner, we replace the goal of imparting information with the goal of learning information about the specified QUD and (2) ...

RDH: any other major differences?

Simulations

Before presenting

Experiment 1: Questions and answers in a hierarchical world

In order to test how questioners choose questions when given a decision problem, and how answerers choose answers under uncertainty about this decision problem, we designed a guessing-game task played by two players: a guesser (the questioner) and a helper (the answerer). In this game, 4 animals (a dalmatian, a poodle, a siamese cat, and a goldfish) were hidden behind 4 gates. Note that these animals were arranged in a class hierarchy (see Fig. 1). The guesser was given a private goal of finding one of the objects, and the helper had privileged information about the exact locations of each object. Before choosing a gate, the guesser could ask the helper a single question, and the helper could respond to this question by revealing the object behind a single gate.

In terms of our model specification, the world space $\mathcal{W}$ is the set of $4! = 24$ possible assignments of four objects to four gates and the goal space $\mathcal{G}$ is the set of four objects that the guesser could possibly be trying to find, and the answer space $\mathcal{A}$ is the set of four gates that the helper could possibly reveal. The key constraint in the task, however, is that the guesser must choose from a *restricted set of questions*: they may be trying to find the goldfish, but cannot directly ask 'where is the goldfish?' Instead, the question space $\mathcal{Q}$ is the set of highlighted nodes in the hierarchy, including higher order nodes that are consistent with multiple answers.

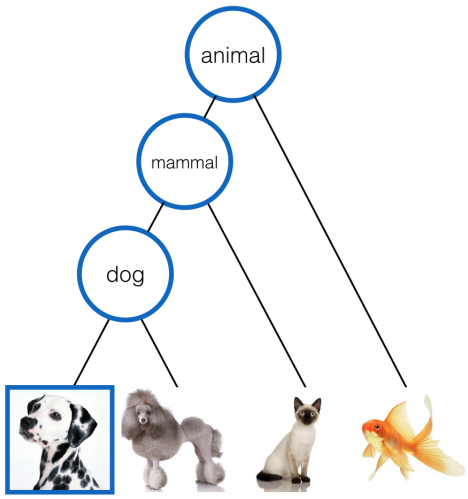

Figure 1: Stimulus hierarchy used in experiment 1. The goal space and answer space contained the four object hidden behind gates (the nodes of the tree). The question space, however, was restricted to the highlighted nodes, proceeding up the hierarchy. This allows us to test the extent to which the answerer infers questioner goals. For example, both the dalmatian and poodle would be truthful responses to the explicit question 'dog?' but only one is the true goal.

RDH: Need to define meaning functions for the interpreter, and actually highlight the nodes in the tree...

We recruited 25 participants from Amazon Mechanical Turk to participate in this task. Each participant provided responses for four trials in the role of the questioner (corresponding to each goal in $\mathcal{G}$), and four trials in the role of the answerer (corresponding to each question in $\mathcal{Q}$). In order to collect responses for all elements of $\mathcal{G}$ and $\mathcal{Q}$, the order of the questioner and answerer blocks was randomly assigned, and the order of stimuli within these blocks was also randomized[1]. Two participants were excluded due to self-reported confusion about the task.

Results for the guesser role are shown with our model predictions in Figure 2(a). There are two primary trends to note in these data. First, questioners tend to choose the indirect question node closest to their goal when the direct question is unavailable. It is unsurprising that they ask about the 'dalmatian' when looking for the dalmatian, but interesting that they strongly prefer asking about a 'dog' when looking for the poodle, about a 'mammal' when looking for the cat, and so on.

[1]All materials are available at https://github.com/hawkrobe/Q_and_A/tree/master/experiment3/versions/experiment3_short

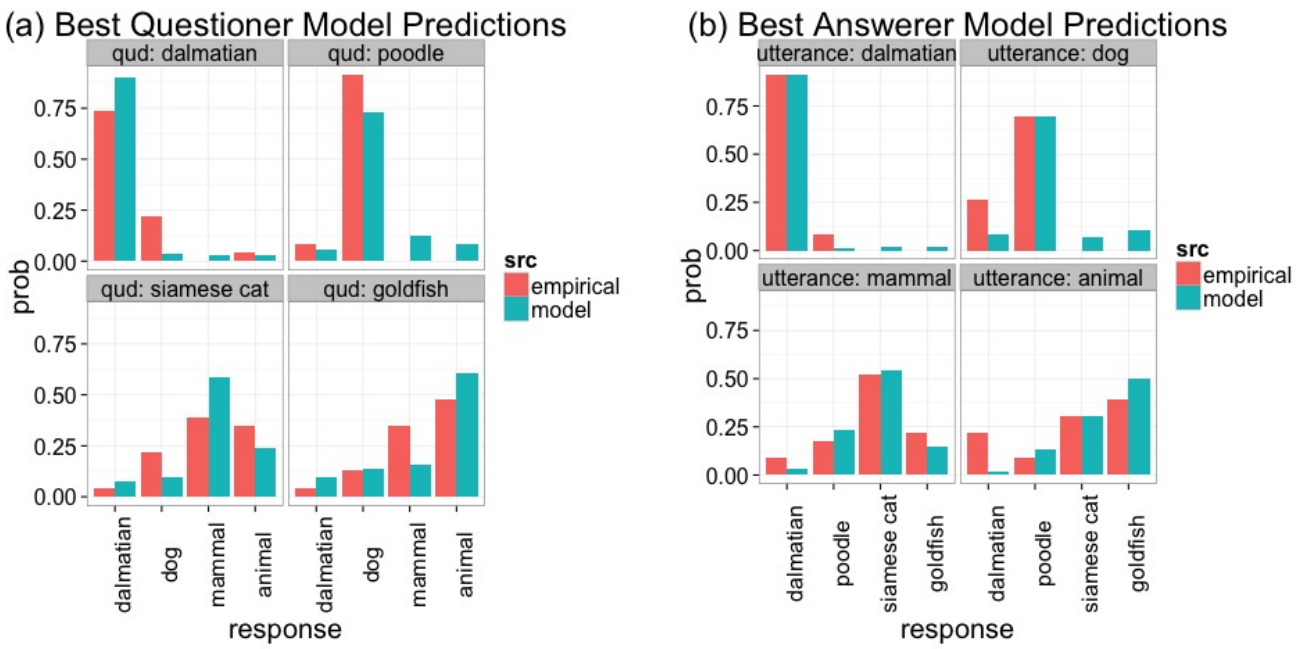

Figure 2: These plots show our results from experiment 1, compared with the predictions of our best-performing models for both questioner and answerer behavior. In both cases, the pragmatic agent performed the best.

RDH: Although, I bet if we actually did a model comparison, the extra .04 correlation you get from going to the pragmatic questioner doesn't justify its complexity... which would be a pretty interesting conclusion, actually

RDH: need to statistically test whether this bar is really bigger than the others, which means I need to rerun the experiment for more power.

Results for the helper role are shown in Figure 2(b). The most striking feature of these data is how closely the answerer distribution matches the questioner distribution

RDH: Possibly just because people did both... It might be the second-timers driving the effect, so we should separate out the different orders, or run it again between-subjects

. If the helper is asked about a 'dog', the dalmatian and poodle would be equally good literal answers to this question, but they strongly prefer to give the location of the poodle. Similarly, the dalmatian, poodle, and siamese cat are all mammals, but helpers prefer to respond to the 'mammal' question by revealing the location of the cat.

We now compare these results to the predictions of our family of models (Fig. 3). At the very simplest level, our *literal answerer* yields a uniform distribution over the four answers that are the case in the given world. This has important consequences for the corresponding *literal questioner* model: when this questioner reasons about which question would generate the most helpful answer from the literal answerer, they find no differences, and therefore have no preference over which question to ask. This is clearly not how questioners and answerers behave and we will not consider the literal models options further.

At a slightly higher level of sophistication, we have an *explicit answerer*, which is equally likely to give the location of all nodes under the explicit class mentioned in the question (e.g. *dog* or *animal*). This model produces the answerer predictions depicted in Figure 3(b), which successfully down-weights the least relevant alternatives, giving a model-data correlation of $r = 0.85$

stats

, but cannot break the symmetries between the explicitly relevant alternatives. For example, it is equally likely to respond 'dalmatian' and 'poodle' when asked about the 'dog', whereas participants strongly preferred 'poodle.' The *pragmatic answerer* model, which uses the question utterance to infer the questioner's underlying goal, is able to break this symmetry

The *explicit questioner*, which reasons about an internal model of an explicit answerer, gets a . When the questioner selects questions that maximize the expected information gain of the explicit answerer's response, it produces the distribution of questions shown in Figure **??**(a) ...

RDH: Need to discuss how we fit the rationality parameters somewhere

Finally, we examine a pragmatic answerer that begins by estimating the likelihood of the questioner having a goal g given the question being asked. It then applies this distribu-

tion of goals as the criterion for whether an answer is useful: it prefers answers for which the distribution of consistent worlds have the item specified by the particular goal is in the right position. The results for this answerer are depicted in Figure **??**(c).

To address concerns that the preceding results were due to particular features of the design such as the one-to-one mapping from goals to questions and from questions to answers, which gives the task the sense of an 'elimination game,' we ran a second experiment using a larger hierarchy ...

Related work

Our account of question and answer behavior ultimately converges on a similar solution as contemporary decision theoretic or game theoretic accounts in linguistics. These theories were a response to early work on question and answer semantics, which focused on the notion of informativeness. In Groenendijk & Stokhof's (1984) theory of question and answer semantics, asking a question induces a partition over the space of possible worlds, where each cell of the partition corresponds to a possible answer. An answer, then, consists of eliminating cells in this partition, and the most useful answers are those that eliminate all relevant alternatives to the true world. However, as van Rooy (Van Rooy, 2003) and others (Ginzburg, 1995) have pointed out, this predicts that *wh*-questions like "Where can I buy an Italian newspaper?" can only be fully resolved by exhaustively mentioning whether or not such a newspaper can be bought at each possible location. Clearly, this is not the case: a single nearby location would suffice. These theories also cannot account for other contextual variation in what counts as a useful answer, such as questions like "where are you?"

More recent theories have tried to fix these problems by introducing some consideration of the questioner's goals. van Rooy (2003), for instance, formalizes these goals as a decision problem faced by the questioner. A useful answer under this decision theoretic account is one that maximizes the expected value of the questioner's decision problem. A useful question is one that induces a sufficiently fine-grained partition, optimally distinguishing the worlds relevant to the decision problem. While this framework elegantly accounts for the context-dependence and relevance-maximization of question and answer behavior, it assumes that the questioner's decision problem is known *a priori* by the answerer. If this were the case, the act of asking questions would seem irrelevant: why wouldn't the answerer directly tell the questioner which action to take?

General discussion

Humans are experts at inferring the intentions of other agents from their actions (Tomasello, Carpenter, Call, Behne, & Moll, 2005). Given simple motion cues, for example, we are able to reliably discern high-level goals such as chasing, fighting, courting, or playing (Barrett, Todd, Miller, & Blythe, 2005; Heider & Simmel, 1944). Experiments in psycholinguistics have shown that this expertise extends to speech acts as well. Behind every question lies some goal or intention. This could be an intention to obtain an explicit piece of information ("Where can I get a newspaper?"), signal some common ground ("Did you see the game last night?"), test the answerer's knowledge ("If I add these numbers together, what do I get?"), politely request the audience to take some action ("Could you pass the salt?"), or just to make open-ended small talk ("How was your weekend?"). These wildly different intentions seem to warrant different kinds of answers, even if the explicit question is expressed using the same words.

In this paper we have presented computational-level evidence that answerer behavior is best described by a pragmatic model that reasons about questioner intentions, using the question utterance as a signal. This analysis provides a novel perspective on the role of questions in dialogues: it is well-accepted under the Gricean view that answerers strive to be relevant, but we find that there is also a burden placed on the questioner to provide sufficient information about should be considered relevant in the first place. By artificially restricting the question space in our experiments, we have shown that when necessary, questioners can reliably choose an optimally informative signal.

XXX

References

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*(4), 313–331.

Boër, S. E., & Lycan, W. G. (1975). Knowing who. *Philosophical Studies*, *28*(5), 299–344.

Clark, H. H. (1979). Responding to indirect speech acts. *Cognitive psychology*, *11*(4), 430–477.

Der Henst, V., Carles, L., & Sperber, D. (2002). Truthfulness and relevance in telling the time. *Mind & Language*, *17*(5), 457–466.

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

Gibbs Jr, R. W., & Bryant, G. A. (2008). Striving for optimal relevance when answering questions. *Cognition*, *106*(1), 345–369.

Ginzburg, J. (1995). Resolving questions, i. *Linguistics and Philosophy*, *18*(5), 459–527.

Goodman, N. D. (2013). The principles and practice of probabilistic programming. In *ACM SIGPLAN Notices* (Vol. 48, pp. 399–402).

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Goodman, N. D., & Stuhlmüller, A. (electronic). *The design and implementation of probabilistic programming languages.* Retrieved 2015/1/16, from `http://dippl.org`

Groenendijk, J., & Stokhof, M. (1984). On the semantics of questions and the pragmatics of answers. *Varieties of*

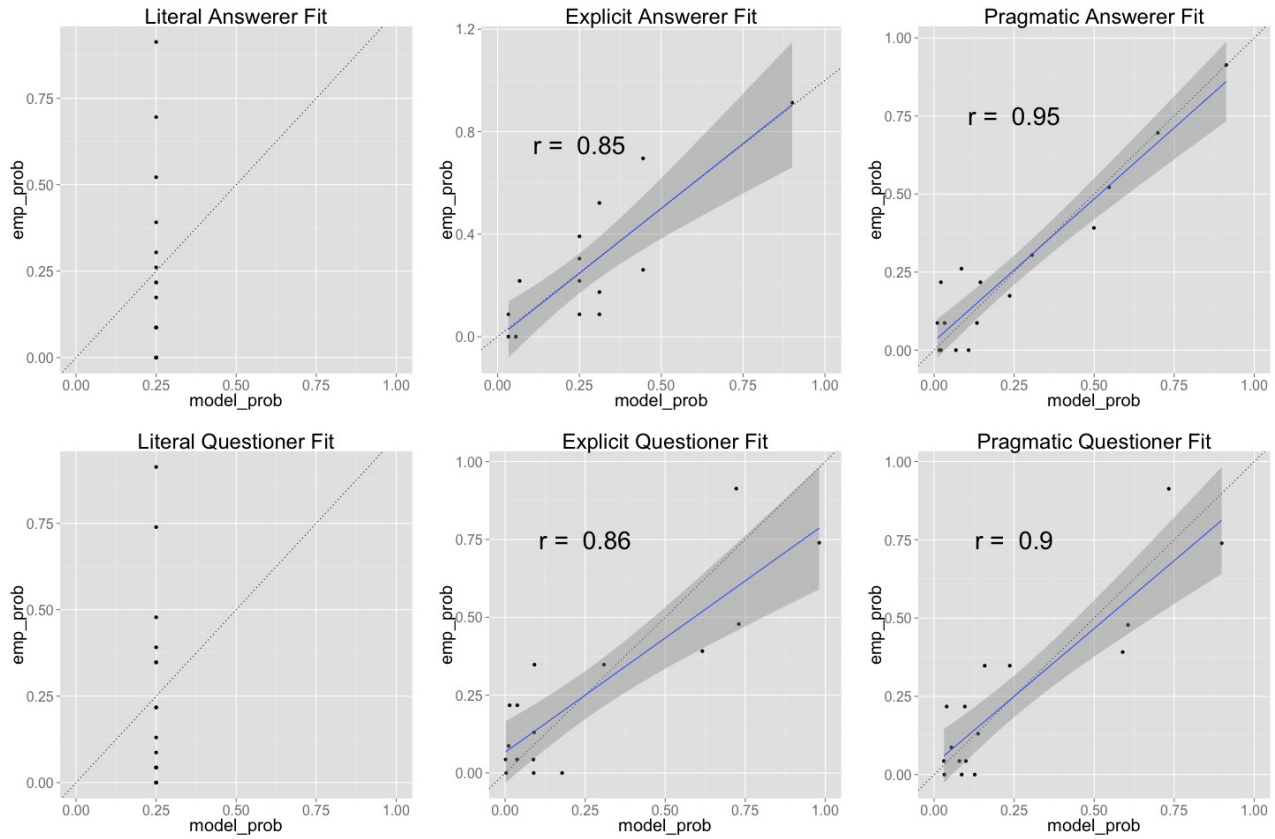

Figure 3: Full space of models, and their correlations with the data from Experiment 1. The questioner models in the second row reason about the answerers directly above them in the top row, and the pragmatic answerer reasons about the explicit questioner. Note that the explicit answerer model was fit using one rationality parameter, the explicit questioner was fit using two optimality parameters, the pragmatic answerer was fit using four parameters, and these parameters were fixed in the pragmatic questioner

RDH: There needs to be more discussion of parameter fitting in the body – the fact that the better models just have more parameters is too easy a target! Maybe if we propagate up the best vals, so that each model only has to fit one or two to its own data?

formal semantics, *3*, 143–170.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 243–259.

Potts, C. (2012). Goal-driven answers in the cards dialogue corpus. In *Proceedings of the 30th west coast conference on formal linguistics* (pp. 1–20).

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, 91–136.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(05), 675–691.

Van Rooy, R. (2003). Questioning to resolve decision problems. *Linguistics and Philosophy*, *26*(6), 727–763.