Teasing apart models of pragmatics using optimal reference game design

Irene Zhou[1] **Jennifer Hu**[1] **Roger P. Levy**[1] **Noga Zaslavsky**[1,2,3]

[1]Department of Brain and Cognitive Sciences, [2]Center for Brains Minds and Machines, [3]McGovern Institute for Brain Research
Massachusetts Institute of Technology
{zhoui, jennhu, rplevy, nogazs}@mit.edu

Abstract

How do humans produce and comprehend language in pragmatic ways? A variety of models of pragmatic inferences have been proposed, and these models are often evaluated on their ability to account for human inferences in reference game experiments. However, these experiments are not tailored to target theoretical differences between models or clearly tease apart model predictions. We propose an optimal experiment design approach to systematically construct reference games that can optimally differentiate between models of human pragmatic reasoning. We demonstrate this approach and apply it to four models that have been debated in the literature: Grammar-based, Iterated Best Response (IBR), Rational Speech Act (RSA), and a recent variant of RSA grounded in Rate–Distortion theory (RD-RSA). Using these optimal reference game experiments, we find empirical evidence favoring iterated rationality models over the grammar-based model, as well as support for the relevance of Rate–Distortion theory to human pragmatic inferences. These results suggest that our optimal reference game design framework may help adjudicate between computational theories of pragmatic reasoning.

Keywords: pragmatics; rational speech act; optimal experiment design; reference games

Introduction

When humans communicate, we convey and interpret meaning beyond what is explicitly said (Grice, 1975; Horn, 1992). The variation and context-dependence of such pragmatic behaviors have made it notoriously challenging to develop a model of pragmatics that can make quantitative predictions.

In the past two decades, there have been a number of proposals formalizing how pragmatic inferences are computed. In particular, a recent debate has centered upon the role of grammar versus game-theoretic reasoning in computing pragmatic implicatures. The Grammatical approach proposes that implicatures are derived entirely within the grammar via a silent operator which essentially negates the meaning of relevant alternatives (Chierchia, 2004; Fox, 2007; Fox & Katzir, 2021; Asherov, Fox, & Katzir, 2021b). In contrast, a class of Iterated Rationality Models (IRMs) formulates speakers and listeners as cooperative agents that reason about each others' beliefs and goals. One well-known instance of an IRM is the Iterated Best Response (IBR) model, which describes agents' optimal strategies under rationality assumptions about their partners (Jäger, 2011; Franke, 2011). Another prominent IRM is the Rational Speech Act model (RSA; Frank & Goodman, 2012; Goodman & Frank, 2016), which builds upon work in Bayesian cognitive modeling (Tenenbaum, Kemp, Griffiths, & Goodman, 2011) as a way of representing uncertainty over possible states of the world. More recently, it has been shown that RSA can be grounded in Shannon's Rate–Distortion theory, yielding another model class called RD-RSA (Zaslavsky, Hu, & Levy, 2020, 2021).

Given this diverse space of proposals, how can we compare different models as candidate theories of pragmatics? The general approach for comparing two models A and B is as follows. Beginning with some shared input and model-specific parameters, each model is run forward to generate a set of predictions in response to the input. The input is used to collect human behavioral data, which is compared against each model's predictions. Typically, the input is hand-crafted to represent a specific phenomenon, such as scalar implicature (e.g., Goodman & Stuhlmüller, 2013; Frank, Emilsson, Peloquin, Goodman, & Potts, 2016), hyperbole (Kao, Wu, Bergen, & Goodman, 2014), or free choice inference (Champollion, Alsop, & Grosu, 2019).

While it is important to compare models' predictions on hand-crafted inputs, this approach faces two main issues in practice. First, different models often make similar quantitative predictions, making it difficult to discriminate between them. Second, it is often unclear how to manually design an experiment that is well-suited for exposing qualitative differences between models. Indeed, while a large body of work has conceptually compared IRMs (e.g., Franke & Jäger, 2014; Franke, 2017) or fit specific IRMs to human data, Benz and Stevens (2018) write "there is no established criterion that would enable an objective comparison" between them. As such, the current empirical pragmatics landscape consists of mixed results. On the one hand, there appears to be a large body of empirical support for RSA (e.g., Goodman & Frank, 2016; Bergen, Levy, & Goodman, 2016), whereas recent studies lend support to grammar-based theories over IRMs (Franke & Bergen, 2020; Asherov, Fox, & Katzir, 2021a; Asherov et al., 2021b). In addition, there have been alternatives proposed to RSA that are theoretically motivated but have yet to be fully evaluated empirically (e.g., Zaslavsky et al., 2020, 2021), further complicating the picture.

In this work, we address these challenges by proposing a framework for optimal reference game design that can be used to tease apart various models of pragmatic reasoning. Specifically, we build on Myung and Pitt's (2009) Optimal

Code and data: github.com/zhouire/pragmatics-oed

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Experiment Design (OED) approach to facilitate a systematic comparison of models of pragmatics. Our approach takes two models of pragmatics, and finds a design for a reference game that yields maximally different predictions. These designs can then be translated into an executable reference game experiment in order to empirically evaluate the two models. We demonstrate this approach for differentiating between the four aforementioned models (Grammatical, IBR, RSA, and RD-RSA). In particular, we first compare across the Grammatical and IRM model classes, and then within the class of IRMs (IBR, RSA, RD-RSA). We find evidence favoring IRMs over the Grammatical model, as well as empirical support for RD-RSA within the IRM class, which further supports the proposal that Rate–Distortion theory may help to explain human pragmatic reasoning (Zaslavsky et al., 2020).

Background: Models of pragmatic reasoning

Before laying out our framework for optimal reference game design, we begin by reviewing the four models of reasoning which we aim to differentiate: the grammatical model, Iterated Best Response (IBR), Rational Speech Act (RSA), and Rate–Distortion Rational Speech Act (RD-RSA).

The grammatical model

Under the grammatical model, pragmatic implicatures are derived within the grammar as part of a sentence's meaning (Chierchia, 2004; Fox, 2007; Fox & Katzir, 2021). The Grammatical approach is executed in two parts: Innocent Exclusion (Fox, 2007) and Innocent Inclusion (Bar-Lev & Fox, 2017). The approach starts with a speaker's assertion S and a set of alternative utterances M (which, in a reference game format, may be inferred by the listener). Innocent Exclusion states that an alternative m can be innocently excluded if m is in all maximal sets of alternatives that can be negated without contradicting the assertion S. After all innocently excludable alternatives have been removed, Innocent Inclusion considers maximal sets of alternatives that can be affirmed consistently with the assertion S, and includes those that appear in all sets.

For simple, single-word utterances in a reference game, the set of alternative utterances is assumed to be the set of single-word utterances U represented in the lexicon, of which one is chosen to be the speaker's assertion S. All alternative utterances $U - S$ can be innocently excluded, and only S can be innocently included. Thus, if a provided referent contains only the asserted feature - and no other features - then the listener will choose that referent. If all referents containing the asserted feature also contain an additional feature, then the listener will be unable to determine which of those referents is the intended meaning, and no implicature arises.[1]

[1]The "no-implicature" output of the Grammatical model is not compatible with a forced-choice paradigm, as in the case of reference games. For consistency with the other models evaluated here, we operationalize the lack of implicature as a uniform distribution over all semantically viable referents. However, see Asherov et al. (2021b) for an alternative method, and Jasbi, Waldon, and Degen (2019) for more general discussion about linking hypotheses.

Iterated Rationality Models (IRMs)

Next, we turn to the class of Iterated Rationality Models (IRMs). In contrast to the Grammatical model, IRMs derive pragmatically enriched meaning "on top of" semantic meaning by formulating speakers and listeners as cooperative agents that reason about each other.

IBR. The Iterated Best Response model (IBR; Jäger (2011)) describes the behavior of rational speakers and listeners arranged under a *cognitive hierarchy* (Camerer, Ho, & Chong, 2004). A level-0 player is constrained only by truthfulness, and a level-$(t+1)$ player acts rationally under the assumption that their partner is a level-t player.

More concretely, suppose there is a set of meanings $\mathcal{M}$ and a set of possible utterances $\mathcal{U}$. In order to convey a meaning $m \in \mathcal{M}$, a level-0 speaker selects uniformly at random from all utterances $u \in \mathcal{U}$ that are literally true of m. For $t > 0$, the level-t players are defined as producing the *best response* with respect to the player at level $t-1$. This strategy assigns equal probabilities to all best responses and zero otherwise:

$$L_{t+1}(m|u) \propto \begin{cases} 1 & \text{if } m = \text{argmax}_{m \in \mathcal{M}} S_t(u|m)P(m) \\ 0 & \text{o.w.} \end{cases} \quad (1)$$

$$S_{t+1}(u|m) \propto \begin{cases} 1 & \text{if } u = \text{argmax}_{u \in \mathcal{U}} L_t(m|u) \\ 0 & \text{o.w.} \end{cases} \quad (2)$$

In this paper, we describe IRM predictions in terms of iterative "depth", which increments by 1 after a full iteration through both speaker and listener has occurred. For consistency, we establish that a level-t IBR player is at depth $\lfloor \frac{t}{2} \rfloor$.

RSA. Similar to IBR, the Rational Speech Act model (RSA; Frank and Goodman (2012)) defines a hierarchy of player types: literal players who are only constrained by truthfulness, and pragmatic players who act rationally by taking their partners into consideration. Unlike IBR, however, RSA players define probability distributions over responses.

The basis of an RSA model is a lexicon function $\mathcal{L}$, which takes an utterance u and meaning m and returns a value in $\{0,1\}$ indicating whether u is literally true of m. We ground the RSA model in a literal speaker S_0, which observes a meaning m and defines a probability distribution over possible utterances u according to the lexicon:

$$S_0(m|u) \propto \mathcal{L}(u,m)\,. \quad (3)$$

Next, RSA defines a pragmatic level-t listener, which is Bayesian with respect to the level-t speaker:

$$L_t(m|u) \propto S_t(u|m)P(m)\,. \quad (4)$$

Finally, RSA defines a pragmatic level-t speaker, which is a distribution over utterances u conditioned on meaning m:

$$S_t(u|m) \propto \exp(\alpha(\log L_{t-1}(m|u) - \kappa(u)))\,. \quad (5)$$

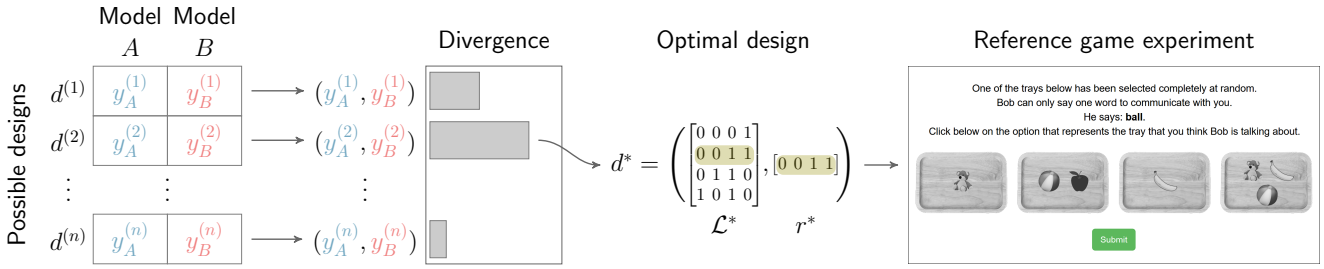

Figure 1: Diagram of our approach, based upon optimal experiment design for model discrimination (Myung & Pitt, 2009). We exhaustively generate all designs and compare the predictions of the two models using a divergence metric D. The design that maximizes D is selected as optimal and automatically translated into an executable reference game experiment.

The level-t speaker maximizes the likelihood that a level-$(t-1)$ listener will recover m upon observing u, and minimizes the cost $\kappa(u)$ of producing u. The sharpness of the distribution is controlled by a free parameter $\alpha > 0$, representing the degree to which the speaker seeks to maximize utility. For RSA, the iterative depth of a player is equivalent to its level.

RD-RSA. Recently, Zaslavsky et al. (2020) showed that with a relatively simple modification, RSA can be grounded in Rate–Distortion theory (Shannon, 1959; Berger, 1971), yielding the RD-RSA model. The listener in RD-RSA is similar to the RSA listener; i.e., it is Bayesian with respect to the speaker. Importantly, however, the pragmatic speaker in RD-RSA differs from the RSA speaker. In RD-RSA, the pragmatic speaker takes into account the marginal probability of producing the utterance u:

$$S_t(u|m) \propto S_t(u)\exp(\alpha(\log L_{t-1}(m|u) - \kappa(u))) \quad (6)$$

$$S_t(u) = \sum_m S_t(u|m)P(m)\,, \quad (7)$$

where $S_t(u)$ is not fixed but rather updated as the speaker reasons about the listener. While RSA and RD-RSA may appear relatively similar, the two models have interesting theoretical differences. For example, they are derived from different optimization principles, and the RSA speaker has a bias toward random utterance production while RD-RSA does not (Zaslavsky et al., 2020).

With these four models in mind, we turn to our proposed framework for optimally designing reference game experiments for teasing apart models of pragmatic reasoning.

Optimal reference game design

Our approach for generating optimal reference games builds on the long tradition of Optimal Experiment Design (OED) in statistics (e.g., Atkinson and Donev (1992)) and psychological research (e.g., McClelland (1997); Holling (2013)). Our goal is to systematically find inputs for which models under comparison are likely to make different predictions. One challenge for comparing models of pragmatics is handling parameters that can freely be adjusted to fit experimental data. Ideally, an experiment should discriminate between two models over a wide range of parameter settings. Therefore, we leverage Myung and Pitt's (2009) OED paradigm for model discrimination, which handles free nuisance parameters via the T-optimality criterion (Atkinson & Donev, 1992; Ponce de Leon & Atkinson, 1991; Uciński & Bogacka, 2005). This criterion is useful for finding designs that maximally discriminate between candidate models of a psychological process.

For the models compared in this study, an input (or "experimental design") consists of two parts: a binary lexicon representing the features and referents in the context, and a row/column of that lexicon corresponding to an utterance/referent used to prompt a listener/speaker. In our study we focus on the *listener* task. Thus, a particular input or design d consists of a lexicon-row pair $(\mathcal{L}, r)$.

Figure 1 illustrates the process of finding an optimal design. We first generate predictions from the models under comparison for each design in the search space. The model prediction for each design d is produced by computing the listener distribution conditioned on row r of lexicon $\mathcal{L}$, which corresponds to a specific utterance. We then evaluate the difference between the model predictions using a divergence metric. The optimal design $d^* = (\mathcal{L}^*, r^*)$ is the one that results in the greatest difference under this measure.

Finding an optimal design

We adapt Myung and Pitt's (2009) approach as follows. Suppose, given a design d and parameters θ_A, model A generates data y_A; in our case, this is the distribution predicted by model A. We can attempt to fit model B to y_A by finding the best-fit parameter vector θ_B^* that minimizes the divergence measure $D_B(d, \theta_A, y_A)$. Because we do not know the parameters θ_A that will best fit human data, we can assess the quality of the design d using the expected value of $D_B(d, \theta_A, y_A)$ assuming a prior distribution over the parameters $P(\theta_A)$ as follows:[2]

$$\int D_B(d, \theta_A, y_A)P(\theta_A)d\theta_A\,. \quad (8)$$

[2]Note that this works even for models with different parameters. For example, RSA and RD-RSA both have a free "rationality" parameter α, unlike IBR and the Grammatical model. We discuss specific parameter priors when we describe each experiment in detail.

This expression quantifies the badness-of-fit of model B conditioned on model A, but we must also take into account the badness-of-fit of model A conditioned on B, resulting in the following global utility function to be maximized, where $P(A)$ and $P(B)$ are model priors (generally uniform):

$$U(d) = P(A)\int D_B(d,\theta_A,y_A)P(\theta_A)d\theta_A + P(B)\int D_A(d,\theta_B,y_B)P(\theta_B)d\theta_B \quad (9)$$

In practice, we evaluate this integral numerically. By fitting each model to predictions made by the other, we assume that the generating model is the true model that captures human behavior, while the fitting model is an impostor model. By maximizing the badness-of-fit of the impostor model, we maximize the likelihood that, given the experimental data, we will select the true model as the better model.

Divergence function. As demonstrated by Equation (9), the divergence function determines what experimental designs emerge as "optimal". For models that output probability distributions (e.g., RSA and RD-RSA), some natural options include distance metrics such as Jensen-Shannon Divergence. In our study, we used maximum rank difference (MRD) as the divergence function. Taking L_A and L_B to be the listener distributions predicted by models A and B, respectively,

$$\begin{cases} \max_m(|L_A(m|u) - L_B(m|u)|) & \text{if } \exists u \text{ s.t. } R(L_A(u)) \neq R(L_B(u)) \\ 0 & \text{o.w.} \end{cases}, \quad (10)$$

where R is the standard competition ranking. MRD is an interpretable measure of lexicon optimality, whereby model predictions are only considered different if the referents are ranked differently in the resulting probability distribution.

Lexicons. We evaluated all valid 4x4 lexicons,[3] defined as binary matrices with no all-0 or all-1 rows/columns, and no duplicate rows/columns. Additionally, we aimed to minimize the effects of experimental noise by only considering designs in which the selected row contained two 1s (matching exactly two meanings), thus giving participants only two reasonable referent choices. In an experimental setting, the lexicon is interpreted indirectly from the visual context; we discuss this in detail in the following section.

From design to data: Reference games

In order to obtain human behavioral data to evaluate the models, we need to translate the output of the OED algorithm (an optimal design) into an executable experiment. Any experimental paradigm that is characterized by a lexicon – or any unique setting for the initial conditions of the model(s) of interest – can be used. Here, we use one-shot reference games, which are a popular experimental tool for studying ad-hoc scalar implicatures (Stiller, Goodman, & Frank, 2015; Frank et al., 2016). A reference game is defined by a set of utterances $\mathcal{U}$ and referents ("meanings") $\mathcal{M}$, and involves two players: a speaker, who attempts to communicate a meaning $m \in \mathcal{M}$ by selecting an utterance $u \in \mathcal{U}$; and a listener, who attempts to recover m upon observing u. The players share the goal of the listener correctly recovering m.

We choose reference games as the experimental paradigm for this study, as a game can be represented by a lexicon matrix. The speaker's task can be represented by a column from this matrix (conditioning on an intended meaning), and the listener's task can be represented by a row (conditioning on an observed utterance). Despite their simplicity, reference games elicit pragmatic behaviors by requiring partners to reason about context to achieve a shared goal.

Experiment 1: Grammatical vs. IRMs

We begin illustrating our approach by comparing across the Grammatical and IRM model classes. As mentioned earlier, recent studies have found empirical support for grammar-based theories over IRMs (Franke & Bergen, 2020; Asherov et al., 2021a, 2021b).

Optimal lexicon. We ran the OED algorithm described above, pairing the Grammatical model with each of the three IRMs. We made several simplifying assumptions to make the integration over parameters tractable: a uniform prior over meanings, no utterance cost, and iterative depth 1.[4] For both RSA and RD-RSA, we set the prior over α to be uniform over the interval $[1,3]$, reflecting minimal expectations about reasonable values of α based on existing empirical studies and the theoretical role of α (Zaslavsky et al., 2020). We leave a detailed treatment of priors to future work.

For each model comparison, we obtained a list of experimental designs ranked from most to least optimal (highest to lowest global utility $U(d)$). After removing the designs that produced a global utility of zero for at least one of the model comparisons, the same design was ranked first for all three comparisons. This optimal lexicon is shown in Figure 2a, with the optimal row (corresponding to utterance u^*) highlighted. Each subplot corresponds to a particular utterance-meaning pair (u^*, m), with model listener predictions $L(m|u^*)$ shown as bars. By design, the Grammatical and IRM listeners make qualitatively differing predictions for u^*. Because no referent contains only the asserted feature ("ball" in Figure 2a), the Grammatical model predicts a listener to be equally likely to choose between the two compatible referents R3 and R4. In contrast, the IRMs predict a preference for R3 over R4 – if a pragmatic speaker had meant R4, they could have unambiguously used the first utterance ("apple" in Figure 2a).

[3] The approach can be applied to lexicons of arbitrary size (including non-square dimensions). However, relatively small lexicons are preferred because they are more straightforwardly translated into reference games, and the search cost is more tractable.

[4] While different cost functions and meaning priors will likely induce differences in model predictions, we chose not to manipulate them because it is unclear how to control them in an experiment.

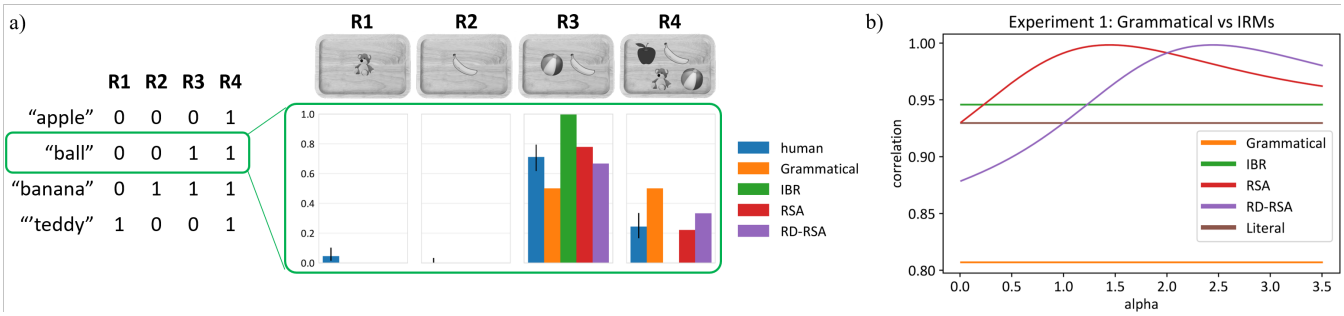

Figure 2: Comparison across Grammatical and IRM classes (Exp. 1). (a) Optimal design selected by OED approach, with human selection rates and model predictions (using $\alpha = 1$ and uniform prior) shown for optimal row. (b) Correlation between model predictions and human data as a function of α (RSA/RD-RSA parameter). Literal represents RSA/RD-RSA at depth 0.

Methods. We used the optimal lexicon and row to conduct a one-shot reference game experiment. Participants were assigned to either the Listener task or the Prior Elicitation task. Following standard methodology (Frank et al., 2016), participants were introduced to a speaker ("Bob") who attempts to describe a target referent using a one-word utterance. On the critical screen, participants were asked to click on the referent that they believed Bob was referring to, based on his utterance (see Reference Game Experiment in Figure 1). In the Listener task, this utterance corresponded to the optimal row identified along with the optimal lexicon, naming a feature present on one or more of the displayed referents. In the Prior Elicitation task, participants were prompted with a masked utterance ("Bob says: **** (you could not hear what he said)"). Afterwards, participants completed a brief exit survey, which contained demographic questions and an attention check asking them to recall the speaker's name.

The referents displayed on the critical screen were constructed using the following method. We first randomly assigned each row of the optimal lexicon to a commonplace object (e.g., apple). Then, we mapped each lexicon column to a "bag of objects" by overlaying images of the objects with a 1 in its corresponding row on top of a neutral tray background (see top of Figure 2a for examples). Object groupings and referent order were randomized. To control for visual salience across referents, all images were displayed in grayscale.

270 participants were recruited via Amazon Mechanical Turk (MTurk) and compensated \$0.20. We restricted this sample to participants with IP addresses in the United States and a 95% approval rating on previous tasks, and prevented users from participating in the study more than once. 25 participants were excluded for incorrectly answering the attention check. The remaining 245 were assigned to either the Listener task (N=111) or the Prior Elicitation task (N=134).

Results. Given u^* (e.g., "ball" in Figure 2a), human participants chose the 2-feature referent (R3, 71.2%) more frequently than the 4-feature referent (R4, 24.3%). We compare the human responses to the predictions made by the Grammatical model and IRMs (RSA, RD-RSA, IBR) at depth 1. For the IRMs, we use the experimentally elicited prior distribution over meanings. Figure 2b shows the correlation (Pearson's ρ) between the model predictions and behavioral data as a function of α. For RSA, the best-fit correlation ($\rho = 0.998$) lies at $\alpha = 1.4$, while for RD-RSA, the best-fit correlation ($\rho = 0.998$) lies at $\alpha = 2.4$. These α values represent a tradeoff that favors maximizing informativeness over minimizing communicative effort, which is consistent with existing empirical investigations of RSA-style models. The similarly high correlations for RSA and RD-RSA is also consistent with prior work suggesting that RD-RSA can account for human behavior as well as RSA (Zaslavsky et al., 2020, 2021). IBR and the Grammatical model achieve lower correlation with the human data ($\rho = 0.946$ and $\rho = 0.807$, respectively). This suggests that probabilistic IRMs (RSA and RD-RSA) are better able to explain the human data than the Grammatical model or IBR, a non-probabilistic IRM.

Experiment 2: Comparing within IRMs

Next, we use our approach to further compare IBR, RSA, and RD-RSA. While Experiment 1 suggests that RSA and RD-RSA may outperform IBR, it is not yet clear whether RSA and RD-RSA can always account similarly for human behavior. Therefore, our approach could be particularly useful for teasing these two models apart.

Optimal input. We ran the OED algorithm described above, comparing RSA vs. RD-RSA, RSA vs. IBR, and RD-RSA vs. IBR. As in Experiment 1, we set the prior over α to be uniform over $[1,3]$, and assumed a uniform prior over meanings, zero utterance cost, and depth 1. Both comparisons produced the same optimal design, shown in Figure 3a. The IRM listeners make qualitatively different predictions for this utterance ("ball"): RSA predicts the listener to prefer R3 over R4, because R4 can be uniquely described with the first utterance ("apple"). However, RD-RSA predicts the listener to prefer R4 over R3, because "ball" has higher marginal probability to be spoken than "apple", which gives it higher weight under the pragmatic speaker and thus increases the pragmatic listener's confidence that the speaker would say "ball" to de-

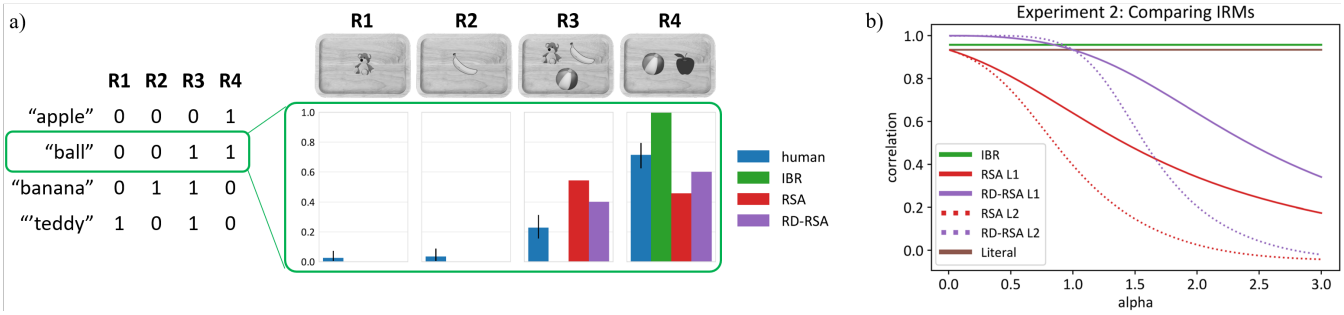

Figure 3: Comparison within the IRM model class (Exp. 2). (a) Optimal design selected by OED approach, with human selection rates and model predictions (using $\alpha = 1$ and uniform prior) shown for optimal row. (b) Correlation between model predictions and human data as a function of α. IBR converges at depth 1. Literal represents RSA/RD-RSA at depth 0.

scribe R3. The IBR listener chooses R4 because a literal speaker is more likely to choose the "ball" out of 2 features, as opposed to the 3 features in R3.

Methods. The experimental methods were identical to those of Experiment 1. 280 participants were recruited, with 18 participants excluded for incorrectly answering the attention check. The remaining 262 were assigned to the Listener task (N=119) or the Prior Elicitation task (N=143).

Results. Given the utterance corresponding to the optimal row of the lexicon (e.g., "ball" in Figure 3a), human participants chose the 2-feature referent (R4, 71.4%) more frequently than the 3-feature referent (R3, 22.7%).

We first compare the distribution of human responses to the predictions made by RSA, RD-RSA, and IBR at iterative depth 1, using the experimentally elicited prior over meanings. Figure 3b shows the correlation (Pearson's ρ) between the model predictions and behavioral data as a function of α. IBR, which does not use the α parameter and converges at depth 1, has a correlation of $\rho = 0.958$ with human data, which lies between the best-fit of RSA and RD-RSA at depth 1. Thus, in contrast to Experiment 1, in this case IBR appears to be comparable with RSA; however, its performance is still below that of RD-RSA. For RSA and RD-RSA, the best fit is achieved at $\alpha = 0.01$ ($\rho = 0.933$) and $\alpha = 0.05$ ($\rho = 1.000$) respectively. These values of α correspond to a strong bias for random utterance production in RSA, and almost entirely non-informative communication in both models (Zaslavsky et al., 2020). This suggests that at least for depth 1, both RSA and RD-RSA predict a non-informative speaker, while attaining high correlation with the human listener data. Interestingly, however, at depth 2 we see an important qualitative difference between RSA and RD-RSA (Figure 3b, dashed lines): RD-RSA achieves maximal performance for a range of αs reaching $\alpha = 1$, whereas the performance of RSA remains high only for values of α near zero.

We further evaluate this using the Fisher Exact Test. The human data is significantly distinct ($p < 0.025$) from IBR and RSA at any α, for depth 1 and 2. In contrast, RD-RSA is not significantly distinct from human data over $\alpha \in [0.01, 0.9]$ at depth 1 and $\alpha \in [0.03, 0.95]$ at depth 2. Therefore, these results lend empirical support to RD-RSA over RSA and IBR.

Discussion

In this work we have proposed a method for optimally designing reference game experiments and used it to tease apart four models of pragmatic reasoning: the Grammatical model and three Iterated Rationality Models (IRMs): IBR, RSA, and RD-RSA. From these experiments, we found evidence favoring IRMs over the Grammatical model, and favoring RD-RSA over RSA and IBR. These findings further support the proposal that Rate-Distortion theory may help to explain human pragmatic reasoning (Zaslavsky et al., 2020), and suggest that our proposed optimal design framework can be used to adjudicate models of pragmatic reasoning.

One limitation of our approach is the computational cost of searching over possible inputs. While an exhaustive search is feasible for low-dimensional binary lexicons, this is generally not tractable. In future work, approximate optimization methods may be used to help improve scalability (e.g., Ouyang, Tessler, Ly, & Goodman, 2018; Foster et al., 2019). Another potential concern is a lack of interpretability of the optimal lexicons identified by the OED approach. The space of possible lexica is so large that an arbitrary lexicon may not correspond to a well-known pragmatic phenomenon. We acknowledge that the OED approach should be seen as a complement to a researcher's judgement.

In future work, the general approach could be leveraged not only to discriminate between established models (as illustrated here), but also to aid researchers in theory-building by identifying points of disagreement between model variants. Indeed, the approach can be applied at several scales of comparison: across model classes (Grammatical vs. IRM), within a model class (IBR vs. RSA vs. RD-RSA), and even within a single framework (e.g., speaker- vs. listener-initialized RSA). This enables an objective comparison of models that are otherwise difficult to tease apart, facilitating a more systematic empirical investigation of models of pragmatics.

Acknowledgments

We would like to thank members of the MIT Computational Psycholinguistics Lab and anonymous reviewers for their helpful feedback and discussion. JH was supported by an NSF Graduate Research Fellowship. NZ was supported by a BCS Fellowship in Computation and a K. Lisa Yang Integrative Computational Neuroscience (ICoN) Postdoctoral Fellowship. RPL acknowledges support from NSF grants BCS-1551866 and BCS-1456081, a Google Faculty Research Award, Elemental Cognition, and the MIT Quest for Intelligence.

References

Asherov, D., Fox, D., & Katzir, R. (2021a). On the irrelevance of contextually given states for the computation of scalar implicatures. In *Linguistic society of america 2021 annual meeting.*

Asherov, D., Fox, D., & Katzir, R. (2021b). *Reference games and the nature of exhaustification.* Retrieved from https://lingbuzz.net/lingbuzz/006257

Atkinson, A. C., & Donev, A. N. (1992). *Optimum Experimental Designs*. Clarendon Press.

Bar-Lev, M. E., & Fox, D. (2017). Universal Free Choice and Innocent Inclusion. In *Proceedings of the 27th Semantics and Linguistic Theory Conference.*

Benz, A., & Stevens, J. (2018). Game-Theoretic Approaches to Pragmatics. *Annual Review of Linguistics*, *4*(1), 173–191.

Bergen, L., Levy, R., & Goodman, N. D. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*.

Berger, T. (1971). *Rate distortion theory; a mathematical basis for data compression*. Prentice-Hall.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004, August). A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics*, *119*(3), 861–898.

Champollion, L., Alsop, A., & Grosu, I. (2019). Free choice disjunction as a rational speech act. In *Proceedings of the 29th Semantics and Linguistic Theory Conference.*

Chierchia, G. (2004). Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In *Structures and Beyond* (pp. 39–103). Oxford University Press.

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T., & Goodman, N. (2019). Variational bayesian optimal experimental design. *Advances in Neural Information Processing Systems*, *32*.

Fox, D. (2007). Free Choice Disjunction and the Theory of Scalar Implicatures. In U. Sauerland & P. Stateva (Eds.), *Presupposition and Implicature in Compositional Semantics* (pp. 71–120). Palgrave Macmillan.

Fox, D., & Katzir, R. (2021). Notes on Iterated Rationality Models of Scalar Implicatures. *Journal of Semantics*.

Frank, M. C., Emilsson, A. G., Peloquin, B., Goodman, N. D., & Potts, C. (2016). *Rational speech act models of pragmatic reasoning in reference games.* Retrieved from psyarxiv.com/f9y6b

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998.

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

Franke, M. (2017). *Game Theory in Pragmatics: Evolution, Rationality, and Reasoning.* Oxford University Press.

Franke, M., & Bergen, L. (2020). Theory-driven statistical modeling for semantics and pragmatics: A case study on grammatically generated implicature readings. *Language*, *96*(2).

Franke, M., & Jäger, G. (2014). Pragmatic Back-and-Forth Reasoning. In *Pragmatics, Semantics and the Case of Scalar Implicatures* (pp. 170–200).

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Grice, H. P. (1975). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics: Speech Acts* (Vol. 3, pp. 41–58). Academic Press.

Holling, H. (2013). Current issues in optimal design. *Zeitschrift für Psychologie*, *221*(3), 121–123.

Horn, L. (1992). The Said and the Unsaid. In C. Barker & D. Dowty (Eds.), *Proceedings of the 2nd Semantics and Linguistic Theory Conference* (Vol. 40, pp. 163–192). Columbus, OH: Linguistic Society of America. doi: 10.3765/salt.v2i0.3039

Jasbi, M., Waldon, B., & Degen, J. (2019). Linking Hypothesis and Number of Response Options Modulate Inferred Scalar Implicature Rate. *Frontiers in Psychology*, *10*, 189.

Jäger, G. (2011). Game Theory in Semantics and Pragmatics. In C. Maienborn, P. Portner, & K. von Heusinger (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (Vol. 3, pp. 2487–2516). Berlin: de Gruyter.

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

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, *2*(1), 3–19.

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*(3), 499–518.

Ouyang, L., Tessler, M. H., Ly, D., & Goodman, N. D. (2018). webppl-oed: A practical optimal experiment design system. In *Proceedings of the Fortieth Annual Conference of the Cognitive Science Society.*

Ponce de Leon, A. C., & Atkinson, A. C. (1991, September). Optimum experimental design for discriminating between two rival models in the presence of prior informa-

tion. *Biometrika*, *78*(3), 601–608.

Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, *4*(142-163), 1.

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc Implicature in Preschool Children. *Language Learning and Development*, *11*(2), 176–190.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279–1285.

Uciński, D., & Bogacka, B. (2005). T-Optimum Designs for Discrimination between Two Multiresponse Dynamic Models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *67*(1), 3–18.

Zaslavsky, N., Hu, J., & Levy, R. (2021). A Rate–Distortion view of human pragmatic reasoning. *Proceedings of the Society for Computation in Linguistics*, *4*. doi: doi.org/10.7275/gc1z-ck09

Zaslavsky, N., Hu, J., & Levy, R. P. (2020). A rate-distortion view of human pragmatic reasoning. *arXiv preprint arXiv:2005.06641*.