# Biases and labeling in iterative pragmatic reasoning

**Jon Scott Stevens (stevens.400@osu.edu)**
Department of Linguistics, The Ohio State University
Columbus, OH 43210 USA

## Abstract

This paper presents a series of reference game experiments (Frank and Goodman, 2012) and fits the results to a number of Bayesian computational models in order to explore the role of linguistic and perceptual bias in iterative pragmatic reasoning. We first discuss the modeling choices made by Franke and Jäger (2016) and others who have used similar frameworks to model reference game tasks. We introduce a space of different plausible Bayesian models based on this work, and compare models' fit to new experimental data to replicate the basic findings of Franke and Jäger (2016) regarding the strong role for perceptual salience (e.g., the primacy of color over shape as a differentiating property for possible referents) and linguistic category (e.g., a preference for nouns over adjectives) in pragmatic reference resolution. We then uncover an additional possible effect of what we call *labeling*, whereby a hearer may simply ignore non-salient, non-differentiating semantic properties, in a manner similar to how an incremental algorithm (Reiter and Dale, 1992) might ignore certain semantic properties when generating referring expressions.

**Keywords:** Iterative pragmatic reasoning; probabilistic pragmatics; reference games; computational modeling; perceptual bias; reference resolution

## Introduction

When someone says, "hold up your finger," you are most likely inclined, without much thought, to hold up your index finger. This may seem unsurprising, as your index finger is particularly salient for a number of reasons. But, as Franke and Degen (2016) point out, further reflection raises the question of why the thumb—which is technically a finger, and which we might expect to be even more salient than the index finger—is never a candidate for reference. The thumb is a prime example of a pragmatic 'blocking' effect: though it is indeed a finger, the existence of the more specific word "thumb" tends to block it from reference by the word "finger". Hence, there is a tension between salience and pragmatic blocking in resolving the referent of "your finger".

This paper presents an exploration of this kind of tension using reference games (Frank and Goodman, 2012; Franke and Degen, 2016, and others). Reference games are communicative tasks where subjects are asked to either produce or interpret short utterances, which are potentially ambiguous in the context, to describe shapes on a screen. Reference games are used as a test of models of iterative pragmatic reasoning, whereby certain potential referents of an utterance are blocked by the existence of a better, more informative alternative utterance available to the speaker.

We further probe work begun in Franke and Jäger (2016) and Stevens (2016) by setting up reference games that favor strong biases toward particular visually salient referents. We test a range of different variants of the Rational Speech Act (RSA) model of Frank and Goodman (2012) on our results. We come to two conclusions:

- We replicate the basic findings of Franke and Jäger (2016), while improving their implementation of RSA by reducing the number of free parameters required from four to one.
- We examine variation between items and uncover a possible effect of what we call *labeling*—an independently motivated mechanism for assigning possibly incomplete semantic labels to potential referents based on salient preferred properties. We show that by introducing labeling into the model, the fit between model predictions and empirical results is improved.

Before diving into these results, we review prior work on reference games, RSA models and bias in reference resolution.

## Prior work

A recent movement toward probabilistic pragmatics—the use of Bayesian, game-theoretic and other similar methods to model how non-literal meaning is conveyed by utterances in context— has been accompanied by an emphasis on using computational models of pragmatic reasoning to explain empirical results (see Franke and Jäger, 2016, for a summary). This includes the rational speech act (RSA) model (Frank and Goodman, 2012; Franke and Jäger, 2016; Bergen et al., 2016, among many others) and its variants, as well as game-theoretic and decision-theoretic models (see e.g. Franke, 2009; Stevens, 2016). These frameworks all tell a similar story at their core: pragmatic phenomena are largely a byproduct of iterated reasoning of the form, 'I expect that she expects that I will say $\phi$ in context C,' or some variant.

**Reference games** A *reference game* task (Frank and Goodman, 2012) is a simple experiment which is designed to elicit iterative pragmatic reasoning behavior. A speaker and a hearer are presented with an array of colored and/or patterned shapes like the one seen in Fig.1. The speaker is assigned one of the shapes and is tasked with choosing a single word to convey to the hearer which shape she has been assigned. For Fig.1 the choices would be "circle," "triangle," "blue" and "red." The hearer receives one of these words from the speaker and tries to guess correctly what was meant. A simple game-theoretic model of Gricean pragmatic reasoning, such as Franke's (2009) iterated best response (IBR) model, makes categorical predictions about the interpretation of ambiguous-in-context words ("triangle" and "blue" in this case). Quite simply, the hearer should assume that the speaker would have used an unambiguous word if she could have, i.e. "red" for the red triangle and "circle" for the blue circle, which leads to the conclusion that either "blue" or "triangle" alone should be taken to refer to the blue triangle. But such categorical models

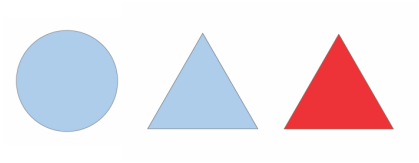

Figure 1: Simple three-image setup for a reference game task.

are typically meant to be normative, and do not aim to reflect the probabilistic nature of how people actually behave. The RSA approach, which builds a Bayesian probabilistic component into a bounded IBR-style reasoning model (Frank and Goodman, 2012; Franke and Jäger, 2016) allows for computational models that more closely match experimental results.

**Rational speech acts** The rational speech act (RSA) approach to modeling pragmatic reasoning computes the probability of a hearer choosing a referent $r$ given a description $d$ via Bayes' rule assuming the speaker chose their utterance *rationally*, which in this case means the speaker attempted to maximize the chance of successful communication. We begin with a function encoding likelihood of referential success of a description $d$ given intended referent $r$ assuming a *naive hearer*—a hearer who randomly selects a referent consistent with $d$'s denotation. Let's call this function $\mathcal{H}_0$.

$$\mathcal{H}_0(r|d) = \frac{1}{|[\![d]\!]|} \text{ if } r \in [\![d]\!], \text{ else } 0 \tag{1}$$

The probability of a rational speaker producing description $d$ to describe intended referent $r$ is taken to be a function of $\mathcal{H}_0$, namely a *soft max* function, which has the effect of approximately maximizing $\mathcal{H}_0$ by introducing a *rationality parameter*, $\lambda$. The rationality parameter encodes the degree to which speakers behave as perfect reasoners. As the value of $\lambda$ increases, this production probability—which we'll call $\mathcal{S}_1$—asymptotically approaches an arg max of $\mathcal{H}_0$. Similarly to Franke and Jäger (2016), we also posit that a *bias function*, $\beta(d,r)$ is added to $\mathcal{H}_0$, which encodes possible prior bias toward certain types of descriptions over others (e.g., a preference for nouns over adjectives, empirically determined for our models). We'll return to the exact nature of the bias function in the next section.

$$\mathcal{S}_1(d|r) = \frac{e^{\lambda \mathcal{H}_0(r|d)+\beta(d,r)}}{\sum_{d'} e^{\lambda \mathcal{H}_0(r|d')+\beta(d',r)}} \tag{2}$$

Finally, the production probability $\mathcal{S}_1(d|r)$ can be plugged in to Bayes' rule, where $P(r)$ is the prior probability of referent $r$ being referred to, to obtain a pragmatically motivated probability function for the hearer, which we will call $\mathcal{H}_2$.

$$\mathcal{H}_2(r|d) = \frac{\mathcal{S}_1(d|r) \times P(r)}{\sum_{r'} \mathcal{S}_1(d|r') \times P(r')} \tag{3}$$

We will use $\mathcal{S}_1$ to make predictions about production probability and $\mathcal{H}_2$ to make predictions about interpretation probability. We use empirically determined values of $P(r)$.

Franke and Degen (2016) also implement a variant of RSA that starts the iteration with the speaker instead of the hearer. That variant begins with a 'literal speaker', which we could call $\mathcal{S}_0$, who randomly selects from among appropriate descriptions. Then $\mathcal{H}_1$ selects referents that maximize the probability of having been referred to by $\mathcal{S}_0$'s utterance, and then a pragmatic speaker $\mathcal{S}_2$ chooses descriptions via Bayes' rule taking $\mathcal{H}_1$ into account. We implement this variant as well.

**Biases and salience** Cognitively oriented pragmatic models like RSA must take into account the prior biases that interlocutors bring to the table. Two such biases factor into Franke and Jäger's model of reference games. Firstly, the authors use data from a prior elicitation task to show that hearers have a prior bias toward picking referents that are more visually salient. For example, there is expected to be a bias toward the red triangle in Fig.1 due to the pop-out effect that its unique color creates. Secondly, the authors use production experiment data to show that there is a prior bias toward using shape nouns rather than color-denoting adjectives to describe an intended object. These biases are built into their probabilistic model, the former being encoded in the prior probability distribution over speaker intentions, and the latter being encoded as the bias parameter $\beta$ which boosts production probability for shape terms. This allows a closer fit to experimental results when compared to more purely Gricean models.

Investigating perceptual bias in the visual domain can shed light on the role of salience in iterative pragmatic reasoning more generally, given that parallels have been found between visual salience and e.g., the use of definite referring expressions (Duan et al., 2013). In this study we find evidence that visual salience affects how hearers assign their own internal semantic labels to the potential referents in a scene. Namely, behavior on certain experimental items suggests that hearers selectively consider properties of potential referents (i.e., the object's color, shape, pattern, etc.) which serve to differentiate them from their competitors. More specifically, we suggest that hearers can generate sets of expected linguistic descriptions for each object using something like an *incremental algorithm* (Reiter and Dale, 1992; Krahmer and Van Deemter, 2012), which has been used to generate referring expressions in a psychologically plausible way. This algorithm is informally sketched in Fig.2. To illustrate, consider the picture in Fig.1. Imagine that the most salient property type is COLOR. To label the red triangle, the algorithm takes its value for COLOR—'red'—and checks whether there is at least one member of the distractor set (the blue triangle and the blue circle) which is not red. There is, and so 'red' gets added to the label set, and both of the non-red items are removed from the distractor set. This leaves an empty distractor set, and so the algorithm halts on the singleton set of labels, {'red'}. The same algorithm will generate {'blue', 'triangle'} for the

- Given an object $O$ in a set of objects $\Omega$, let $L$ be $O$'s *label*—a set of semantic properties (e.g., {'red', 'triangle'}) to describe $O$. Let $P^*$ be an ordered sequence of property types which are ordered by salience (e.g. $\langle \text{COLOR}, \text{SHAPE} \rangle$, if color is more salient than shape). Let $D$ be the set of *distractors*, i.e., $\Omega \setminus O$
- Initialize $L$ to $\{\}$
- For $P$ in $P^*$:

  1. Let $V$ be the value that $O$ has for property $P$
  2. Let $\Omega_{\neg V}$ be the set of objects that have a different value for $P$ than $V$
  3. If $D \cap \Omega_{\neg V} \neq \{\}$, then add $V$ to $L$ and remove all members of $\Omega_{\neg V}$ from $D$
  4. If $D = \{\}$, return $L$

Figure 2: An informal presentation of an incremental algorithm for generating salient and informative semantic labels for referents.

blue triangle and {'blue', 'circle'} for the blue circle. The red triangle is simply labeled as the red thing, because 'red' is a preferred property that uniquely differentiates it, while the other two shapes are labeled according to both color and shape.

## Computational models

We implement a variety of models centered around the RSA implementation of Franke and Jäger (2016), though we reduce the number of free parameters from four to one. The first reduction comes from the choice to use a single value of the rationality parameter $\lambda$ to predict both speaker and hearer behavior, where Franke and Jäger (2016) fit two $\lambda$ values separately. The second reduction comes from the use of empirical data to determine values of $\beta(d,r)$—the observed bias toward nouns in production—where Franke and Jäger (2016) use a pair of fixed values that were tweaked for best performance. Using a speaker norming task, as described in the next section, we obtain a prior probability of noun vs. adjective for each experimental item we want to model. We then set $\beta(d,r)$ to be proportional to this prior probability.

$$\beta(d,r) = \frac{P(d)}{\sum_{d' | r \in [\![d']\!]} P(d')} \text{ if } r \in [\![d]\!], \text{ else } 0 \qquad (4)$$

We now have a single-parameter RSA model that will make predictions about both production and interpretation rates in a reference game task, taking biases into account.

We implement a number of variants of this model to allow us to assess some of the modeling choices we and others have made. In particular, we want to answer the following three questions about our modeling choices:

1. Bias vs. no bias: Do we really need the $\beta(d,r)$ term?
2. Naive hearer vs. literal speaker: Should we really start with a naive hearer $\mathcal{H}_0$, as opposed to with a literal speaker $\mathcal{S}_0$ à la the variant in Franke and Degen (2016)?
3. Uniform level-0 prior vs. empirical level-0 prior: Should the naive hearer and/or literal speaker select randomly from among semantically appropriate actions, as opposed to selecting proportionally to the empirically determined prior?

| | Uniform level 0 | Empirical level 0 |
|---|---|---|
| $\mathcal{S}_1$ / $\mathcal{H}_2$, no bias | F&G (2012) | |
| $\mathcal{S}_1$ / $\mathcal{H}_2$, $\mathcal{S}_1$ bias | F&J (2016) | |
| $\mathcal{S}_2$ / $\mathcal{H}_1$, no bias | F&D (2016) | |
| $\mathcal{S}_2$ / $\mathcal{H}_1$, $\mathcal{H}_1$ bias | | |

Table 1: Eight possible model variants based on the three questions posed, where three of the cells are occupied by examples of a model of that type—Frank and Goodman (2012), Franke and Jäger (2016) and Franke and Degen (2016).

There are a total of $2^3 = 8$ possible combinations of yes/no answers to these three questions, each corresponding to a different model variant. The variant we have described, based on Franke and Jäger (2016), is the "yes/yes/yes" model. That means that bias is implemented, we follow the trajectory $\mathcal{H}_0 \rightarrow \mathcal{S}_1 \rightarrow \mathcal{H}_2$ rather than starting with a non-rational speaker, and $\mathcal{H}_0$ chooses a random semantically compatible meaning, ignoring the prior probability $P(r)$. Table 1 lays out the possibilities and points to examples of a few of the models from the RSA literature.

We implemented all models with integer $\lambda$ values between one and ten[1] and used root mean square error (RMSE) as a measure of overall difference between model predictions and experimentally determined values.

## Experiments

**Participants and materials** We conducted four experiments via Amazon Mechanical Turk, two experiments to determine prior probabilities for referents and descriptions, and two reference game tasks, one where the Turker played the part of the speaker and one where they played the part of the hearer. For each experiment, 100 Turkers were assigned to one of two lists containing nine experimental items, for a total of 18 items. Each item was an array of three images similar to Fig.1, where one image was distinguished from the other two by its shape, another image was distinguished by another salient attribute, and the third image was not distinguished along any dimension. The order of item presentation was randomized, as was the order in which the shapes were presented on the screen. Items fell into one of three categories based on which salient distinguishing attribute was used, with 6 items in each category:

1. Color: Red vs. blue, as in Fig.1
2. Pattern: Striped vs. solid (one striped and two solid)
3. Size: One shape bigger than the other two

Native language was assessed as part of a post-task questionnaire. Subjects were paid $0.70 for about 5 minutes of their time. If any subject's responses were incomplete, or if the

[1]The effect of $\lambda$ on model performance is gradual enough, and the differences between the different model variants large enough, that not much fine-tuning is required to make our point.

subject was not a native speaker of English, the data from that subject was excluded from analysis.

**Experiment 1: Eliciting hearer priors** Following Frank and Goodman (2012) and others, we use empirically determined values for the prior probabilities in our model. The prior probability $P(r)$ of choosing a referent $r$ is taken to be a measure of the salience or 'newsworthiness' of a referent, i.e., a general measure of how likely $r$ is to be talked about. To elicit this experimentally, we asked subjects to select an image to describe, giving them no guidance on which images to select, and then type a description of it. The point of this experiment was not what the descriptions were, but rather which shapes they chose to talk about. We took this as a proxy for the salience of the referent, and thus its prior probability of being referred to. We asked them to type descriptions as a secondary task in order to situate the shape selection within a natural communicative setting. We used data from 97 subjects after exclusions.

For the color items we obtained similar results to Franke and Jäger (2016), where the red shape was picked much more often (probability 0.5) than either of the blue shapes, and where the distinguished blue shape (e.g., the circle in Fig.1) was picked more often (0.33) than the non-distinguished blue shape (0.17). For the size items, we found that the distinguished smaller shape had a high prior probability (0.5 vs. 0.26 and 0.24 for the large and small competitors, respectively), and for the pattern items, the priors were closer to equal for the striped and distinguished solid shapes (0.36 and 0.40, respectively), and lowest for the non-distinguished shape (0.24).

**Experiment 2: Eliciting speaker priors** To empirically determine whether and to what extent speakers are biased toward nouns like 'circle' over adjectives like 'red', we ran an experiment just like Experiment 1, but with two important differences: (i) subjects were assigned one of the three images to describe, rather than being asked to pick one themselves (image assignments were counterbalanced across lists so that shape-distinguished, attribute-distinguished and non-distinguished items were equally represented), and (ii) subjects were told to limit their descriptions to a single word, in order to bring the task more in line with a reference game task. To discourage any kind of pragmatic reasoning, subjects were asked to use the 'first word that came to mind' and not to overthink it. We analyzed data from 84 subjects after exclusions, and only looked at items where either a shape-denoting noun or relevant attribute-denoting adjective was used (very few did not fall into this category). The words were input as free text, and thus we hand-tokenized the responses to account for spelling mistakes and superficial lexical differences (e.g., 'big' vs. 'large'). We found an overwhelming prior bias toward nouns.Overall, shape terms were used two-thirds of the time. There is evidence that this task successfully elicited prior linguistic biases and limited the amount of pragmatic reasoning being used to determine descriptions. For example, for the items where an attribute term would uniquely distinguish the intended referent, shape terms nonetheless comprised 60% of responses, more than double the shape-term response rate for Experiment 3, which was designed to elicit pragmatic reasoning.

| | Word | | | |
|---|---|---|---|---|
| Image | $\text{ATTR}_D$ | $\text{ATTR}_N$ | $\text{SHAPE}_N$ | $\text{SHAPE}_D$ |
| $\text{ATTR}_D$ $\text{SHAPE}_N$ | .75 / **1** | .00 / **.00** | .25 / **.52** | .00 / **0** |
| $\text{ATTR}_N$ $\text{SHAPE}_N$ | .00 / **0** | .29 / **.74** | .71 / **.48** | .00 / **0** |
| $\text{ATTR}_N$ $\text{SHAPE}_D$ | .00 / **0** | .03 / **.26** | .00 / **.00** | .97 / **1** |

Table 2: Production of $d$ given $r$ (on the left, sum horizontally to 1) / selection of $r$ given $d$ (on the right in bold, sum vertically to 1) in Experiments 3 and 4. Subscripts $D$ and $N$ mean 'distinguishing' and 'non-distinguishing', respectively, and ATTR stands for 'attribute'.

**Experiment 3: Reference game, speaker role** Experiments 3 and 4 instantiate the canonical reference game task described in the second section. Experiment 3 asks subjects to play the role of the speaker in a reference game. Similarly to Experiment 2, subjects are assigned one of the three images and asked to give a one-word description. But for this experiment, they are explicitly told to select from a list of the relevant words (e.g., "red", "blue", "triangle", "circle"). And unlike Experiment 2, the task is framed as a game. Subjects are told they are sending a message, and to assume that a "receiver" will receive these descriptions and make a guess as to which image was assigned. The goal, they are told, is for the receiver to guess correctly as often as possible. Data from 79 subjects was used.

**Experiment 4: Reference game, hearer role** Experiment 4 asks subjects to play the role of the hearer, or the "receiver". For each item, a single word is displayed at the top of the screen, which the subjects are told has been carefully selected and sent to them by a sender who wants them to correctly guess an image from a one-word description. Word selection was counterbalanced across both lists. The subjects were required to select a single image for each item. Data from 91 subjects was used.

**Results** Our reference game experiment results are in line with other reference game results in the literature, and are summarized in Table 2. Like Franke and Jäger (2016), we find that the expected propensity toward interpreting ambiguous shape and attribute words (like "triangle" and "blue" in Fig.1) as referring to the non-distinguished shape (like the blue triangle) is dampened for the shape words, likely reflecting hearer knowledge of speakers' prior noun bias, where the prior noun bias makes a shape term like "triangle" a less reliable signal that the non-distinguished referent is intended.

| | Uniform level 0 | Empirical level 0 |
|---|---|---|
| $\mathcal{S}_1$ / $\mathcal{H}_2$, no bias | .12 / .20 | .14 / .23 |
| $\mathcal{S}_1$ / $\mathcal{H}_2$, $\mathcal{S}_1$ bias | **.06 / .18** | .09 / .22 |
| $\mathcal{S}_2$ / $\mathcal{H}_1$, no bias | .23/ .20 | .23 / .20 |
| $\mathcal{S}_2$ / $\mathcal{H}_1$, $\mathcal{H}_1$ bias | .25 / .22 | .24 / .23 |

Table 3: RMSE for speaker predictions (left) / hearer predictions (right). Best-case $\lambda$ value used for each reported RMSE. Best model results are in bold.

Table 3 shows how our model variants line up with the empirical results in terms of the root mean square error (RMSE), which is a measure of overall difference between predicted and observed values obtained by calculating the mean of the square of the difference between each predicted vs. observed value and taking the square root. We used the difference in predicted vs. observed subject means for each experimental item (i.e., each array of images) to determine RMSE. Not only is our refinement of Franke and Jäger (2016) the best model to predict these data, but our best-case value of the $\lambda$ parameters ($\lambda = 4$) is the best-case value for both the speaker and hearer model independently. That is to say, we would not do a lot better by allowing for separate $\lambda$ values for speaker and hearer. We take this to be a nice replication of the basic finding of Franke and Jäger (2016), obtained using only one free parameter that was only broadly tweaked.[2]

The numbers in Table 2 are somewhat closely replicated, with every value being within three percentage points of the real value. A plot of predicted vs. actual results from Table 2 is given in Fig.3. However, the numbers in Table 2 are averaged over all items, and tell us nothing about the range of variation of responses for different kinds of images. RMSE gives us an overall assessment of error taking into account error at the level of each individual item. What the RMSE values in Table 3 tell us is that the speaker model fits considerably better than the hearer model.

Why is the hearer model so noisy? Given the proximity of predicted to actual results on average in Fig.3, the source of the noisiness must be coming from differences between item types. An item-level investigation of the source of the higher-than-expected RMSE will lead us to posit that when there are highly perceptually salient options, as in these experiments, hearers are inclined to label their options in a way that is similar to the output of an incremental algorithm (Fig.2).

## Labeling

We now break down by-item behavior further, looking not only at whether the image array was shape- color- or size-distinguished, but also at which word was sent to the hearer. We find that the predicted qualitative pattern—that ambiguous descriptions (and only ambiguous descriptions) should prompt a plurality of guesses of the non-distinguished

[2]Franke and Degen (2016) also consider the combination $\mathcal{S}_1$ / $\mathcal{H}_1$, i.e., a non-iterative model. This would not do any better here, as we see in Table 3 that $\mathcal{H}_1$ never makes better predictions than $\mathcal{H}_2$.

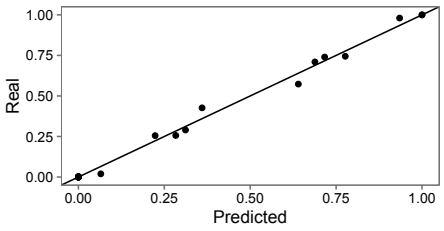


Figure 3: Hearer predictions vs. observed values, averaged over subjects and items.

image—holds in all but two cases. These two cases are depicted in Fig.4, where we see a deviation for (i) color items when the hearer is sent an ambiguous color term and (ii) size items when the hearer is sent an ambiguous shape term. There is a pattern to these deviations. The pattern is that we see a shift away from the non-distinguished image only in cases where the semantically ruled out referent (e.g., the red thing if the description is "blue") has a high prior (in both cases, $\sim$50%). Let's break down what this means for the three item types. First, when the hearer receives "blue" for an item like Fig.1, we find higher-than-expected selection of the unique shape (the circle in Fig.1). This is the item type for which the attribute-distinguished image (the red triangle) is maximally salient according to Experiment 1. Second, when a hearer receives an ambiguous shape term for a size-distinguished item, we find higher-than-expected selection of the uniquely large referent. This is the item type for which the shape-distinguished image is maximally salient according Experiment 1. Finally, the pattern-distinguished items fall entirely in line with what we expect, and those are the items where the priors for shape- and attribute-distinguished images are much closer to each other.

Qualitatively speaking, we would expect this if the referents were labeled according to salient distinguishing properties along the lines of Fig.2, a well-established algorithm for generating referring expressions, which we adapt for generating hearer-internal labels for possible referents. Consider Fig.1 one more time: for the $\sim$50% of subjects in Experiment 1 who chose the red triangle, we can assume that COLOR would be their primary salient property type for purposes of Fig.2. This would generate the labels {'blue circle', 'blue triangle', 'red'}. Assuming these same priors for Experiment 4 (as we have been) we could posit that on $\sim$50% of trials, the subject has this same labeling. In that case, upon hearing the description "blue", the subject would be at chance between the two blue shapes, because under this labeling, the speaker could have used 'triangle' to uniquely describe the blue triangle and 'circle' to uniquely describe the blue circle, leaving no principled way to interpret "blue" other than to guess.

Labeling could explain the qualitative deviations, and even though the numbers are not perfect, it does indeed improve model fit to add a labeling component to the model. We can do this by substituting a new $\mathcal{S}_1$ function $\mathcal{S}_1'$ into the $\mathcal{H}_2$ equa-

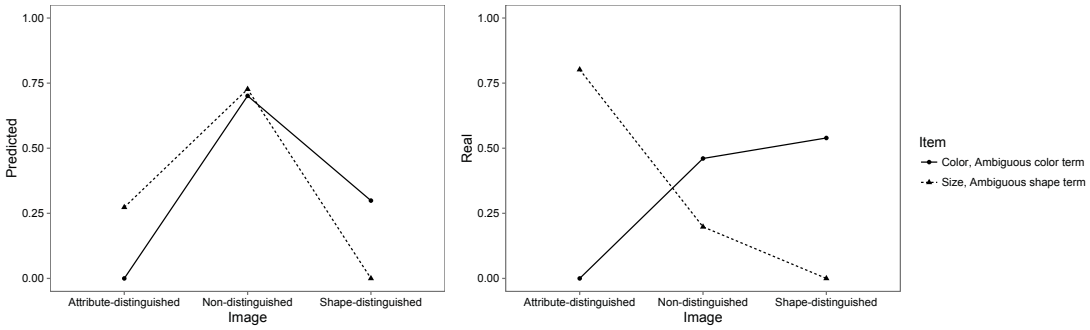


Figure 4: Predicted (left) and actual (right) referent selection for two combinations of item type and description type.

tion which takes labeling into account. Letting $\mathcal{L}$ be the set of labels for each possible referent, $\mathcal{S}_1'$ can be defined as follows:

$$\mathcal{S}_1'(d|r) = \sum_{\mathcal{L}} P(\mathcal{L}) \times \frac{e^{\lambda \mathcal{H}_0(r|d,\mathcal{L})+\beta(d,r)}}{\sum_{d'} e^{\lambda \mathcal{H}_0(r|d',\mathcal{L})+\beta(d',r)}} \quad (5)$$

We introduce no new free parameters if we simply take $P(\mathcal{L})$ to be the prior probability of the shape-distinguished referent for the $\mathcal{L}$ obtained when shape is primary, the prior of the attribute-distinguished referent for the $\mathcal{L}$ obtained when attribute is primary, and the prior of the non-distinguished shape for the 'full' $\mathcal{L}$, which omits no information. Doing this, we can reduce the RMSE from 0.18 to 0.15, and could perhaps reduce it further if we could independently assess how primary salient properties are chosen. Using multinomial choice probabilities to determine log likelihood, we can show that the data from Experiment 4 are significantly more likely under the model with labeling.[3]

## Conclusion

Using a variety of models and experimental items and tasks, we have replicated existing results regarding behavior in reference games, and potentially found a new one, an effect of labeling under conditions where certain referents have highly salient properties. That it might matter how people internally label possible referents is not really a new idea, and is in fact in line with game-theoretic literature on coordination (see e.g., Sugden, 1995). But it provides somewhat of a paradox. On the one hand, this and other studies find that speakers exhibit a bias toward noun descriptions in reference games, across the board, and yet it seems as if hearers are assigning labels to potential referents that in some cases would lead them to expect the opposite (e.g., to expect "red" to describe the red triangle in Fig.1). Thus further work is warranted to probe whether such a mismatch between speaker behavior and hearer expectations is generally observable.

Further work must also be done to probe the details of exactly how labeling works in reference games, and what the implications are for iterative pragmatic reasoning more generally. For example, it remains to be seen whether labeling should be seen as part of a rational process of pragmatic reasoning, or as something that competes with it, as Stevens (2016) would suggest. Finally, future work will use online measures to probe the mechanisms that give rise to the probabilities in our models. This would take us beyond computational-level models, using such models only as a starting point to guide us toward a more fine-grained understanding of this behavior (see Yang, to appear).

## Acknowledgments

This work was funded by the American Council of Learned Societies (ACLS). Thanks to Marie-Catherine de Marneffe, Micha Elsner and the OSU psycholinguistics lab group.

## References

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

Duan, M., Elsner, M., and de Marneffe, M.-C. (2013). Visual and linguistic predictors for the definiteness of referring expressions. In *Proceedings of the 17th SemDial Workshop, Amsterdam*.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. PhD thesis, Universiteit van Amsterdam.

Franke, M. and Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, 11(5).

Franke, M. and Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1).

Krahmer, E. and Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Reiter, E. and Dale, R. (1992). A fast algorithm for the generation of referring expressions. In *Proceedings of the 14th conference on Computational Linguistics, Volume 1*, pages 232–238. ACL.

Stevens, J. (2016). When do we think strategically? *Zeitschrift für Sprachwissenschaft*, 35(1).

Sugden, R. (1995). A theory of focal points. *The Economic Journal*, 105:533–550.

Yang, C. (To appear). Rage against the machine: Evaluation metrics in the 21st century. *Language Acquisition*.

[3]Change in deviance between the two models, $\Delta D = 59.26$, where deviance is -2 times log likelihood, follows a chi-square distribution with degrees of freedom equal to the number of parameters added to the more complex model. Six prior values must be specified to the labeling model—two each for color, shape and size items—yielding $\chi^2 = 59.26,\ df = 6,\ p < 0.001$.