Speak before you listen: Pragmatic reasoning in multi-trial language games

Les Sikos (sikos@coli.uni-saarland.de)[1]

Noortje J. Venhuizen (noortjev@coli.uni-saarland.de)[1]

Heiner Drenhaus (drenhaus@coli.uni-saarland.de)[1]

Matthew W. Crocker (crocker@coli.uni-saarland.de)[1]

[1]Department of Language Science and Technology, Saarland University, 66123 Saarbrücken, Germany

Abstract

Rational Speech Act theory (Frank & Goodman, 2012) has been successfully applied in numerous communicative settings, including studies using one-shot web-based language games. Several follow-up studies of the latter, however, suggest that listeners may not behave as pragmatically as originally suggested in those tasks. We investigate whether, in such reference games, listeners' pragmatic reasoning about an informative speaker is improved by greater exposure to the task, and/or prior experience with being a speaker in this task. While we find limited evidence to suggest that increased exposure to the task results in more pragmatic responses, listeners do show increased pragmatic reasoning after first playing the role of the speaker. Moreover, we find that only in the Speaker-first condition, participant's tendency to be an informative speaker predicts their degree of pragmatic behavior as a listener. Taken together, these findings demonstrate that, in these settings, experience as a speaker enhances the ability of listeners to reason pragmatically about informative speakers, as modeled by RSA.

Keywords: Pragmatic inference; referential expressions; language production; language comprehension; RSA

Introduction

Language is often ambiguous, and during natural communication listeners need to resolve this ambiguity in order to understand what a speaker is saying. Depending on the situation, this may require reasoning about the speaker's perspective, goals, and intentions. The Rational Speech Act (RSA) model, introduced by Frank and Goodman (2012), is a probabilistic Bayesian model that aims to formalize the pragmatic reasoning underlying listener's choices. This model assumes that rational listeners reason about the decision-making of informative speakers, who in turn reason about rational listeners. More specifically, RSA uses Bayesian inference to derive the listener's optimal interpretation of a given utterance, based on the likelihood that speakers would choose that particular utterance to convey the intended message, combined with the prior probability of the interpretation. Frank and Goodman (2012) empirically tested RSA's predictions against human behavior using a one-shot web-based version of the referential communication game (Wittgenstein, 1953), in which each participant saw only a single trial. Figure 1 shows an example of the type of visual contexts used in such games.

Figure 1: Example of a critical visual context in the current study.

The goal of a listener is to select one of the referents based on a given one-word utterance—in this case, a shape word (e.g., "boot") or a color word (e.g., "green"). Importantly, an ambiguous expression like "boot" (or "blue") may trigger listeners to reason about the speaker's choice of word: why use the ambiguous word, if not to refer to the object that cannot be described using any unique features (i.e., the *blue boot*)?

When applied to such a game, RSA determines the probability with which listeners will select a referent given a particular word, based on the probability that speakers would use that word, combined with the prior probability of referring to an object, which is usually determined empirically via a *Salience* task (Frank & Goodman, 2012). That is, the probability that a listener selects referent r_S given word w in visual context C is defined to be proportional to the *informativity* of word w—based on a model of a rational speaker—multiplied by the prior probability of referring to referent r_S.

$$P(r_S|w,C) \propto P(w|r_S,C)P(r_S) \quad (1)$$

Indeed, given the word "boot" in the context shown in Figure 1, RSA predicts that a pragmatic listener will infer that the intended referent is the *blue boot*—assuming a more or less uniform prior—because for the alternative referent (*green boot*) there exists a more informative expression ("green"). In contrast, a non-pragmatic 'literal' listener is predicted to simply rely on the prior probability of the two referents that match the literal interpretation of the given word (i.e., the *blue boot* and the *green boot*). Frank and Goodman (2012) showed that RSA's predictions were strongly correlated with human judgments, which suggests that listeners use pragmatic reasoning in such *one-shot* games despite the artificial, highly constrained, and minimally interactive nature of the task.

However, while the RSA framework has been successfully applied in other domains (Bergen, Levy, & Goodman, 2016; Degen, Hawkins, Graf, Kreiss, & Goodman, 2020; Kao, Wu, Bergen, & Goodman, 2014; Scontras & Goodman,

2017), follow-up studies using one-shot web-based language games suggest that listeners may not behave as pragmatically as previously assumed in these settings (e.g., Frank, Emilsson, Peloquin, Goodman, & Potts, 2016; Franke & Degen, 2016; Qing & Franke, 2015; Sikos, Venhuizen, Drenhaus, & Crocker, 2021). For instance, Qing and Franke (2015) attempted a close replication of Frank and Goodman (2012) but presented participants with only the critical *pragmatically solvable* context types (i.e. those which require the type of pragmatic reasoning illustrated in Figure 1) and found that the basic RSA model was a poor predictor of their data. Moreover, across three experiments Sikos et al. (2021) observed only modest evidence of pragmatic behavior in one-shot reference games: listeners only preferred the pragmatic referent (e.g.. the blue boot in Figure 1) when they observed a color word ("blue"); in contrast, when they observed a shape word ("boot"), listeners instead had a strong preference for the shape competitor (the green boot), which is highly salient due to its unique color. Sikos et al. (2021) further showed that RSA was outperformed by a baseline literal listener model that was driven simply by literal word meaning and the prior probability of referring to an object. This suggests that the pragmatic component within the RSA model (i.e., the speaker model) does not contribute substantially to the high correlation of RSA with listener behavior. This result may at least partially be due to the relatively artificial nature of the communicative task and setting of one-shot web-based language games, and the fact that listeners do not have the opportunity to become familiar with this task.

The goal of the current study, therefore, is to investigate whether the modest evidence of pragmatic behavior in one-shot referential communication games is due to the limited exposure that participants receive in such settings. That is, listeners may not fully recognize that it would be useful to reason about an informative speaker because they only encounter a single trial. Consistent with this possibility, Franke and Degen (2016), Experiment 1, found that listeners (on average) did behave pragmatically in a multi-trial version of the reference game, in which participants played 66 trials in the role of listener, and additionally played the role of speaker in four practice trials before beginning the actual experiment. In order to independently evaluate the influence of these factors on listeners' pragmatic reasoning (as formalised by RSA), we investigate here whether listener behavior is influenced by: (a) greater exposure to the task (i.e. one versus multiple trials), and/or (b) experience with being a speaker in this setting (i.e., Listener-first versus Speaker-first). In order to separate the influence of these factors, we use a block design rather than interleaving speaker and listener trials. In addition to assessing listener behavior, we also test whether the salience of individual referents and Speaker Rationality (i.e. how informative speakers actually are) are modulated by these factors.

We evaluate whether these factors reliably influence pragmatic listener behavior and speaker rationality in two complementary ways. First, we directly compare the behavioral results across conditions. If increased exposure to the task matters, we should see increased pragmatic behavior of participants after exposure to multiple trials. If having experience as the speaker matters, we should see more pragmatic behavior in the Speaker-first condition relative to the Listener-first condition. Second, we compare the observed listener behavior to the predictions of two Bayesian models: (1) RSA, for which we directly incorporate the observed speaker behavior as the likelihood of using particular words, as well as the observed salience judgments (described below) as estimates for the prior probability, and (2) a baseline Literal Listener model that that does not assume pragmatic reasoning (following Sikos et al., 2021). If RSA provides a better fit than the baseline model in one of the conditions (e.g., in the Speaker-first condition), it would provide evidence that listeners behave pragmatically—in terms of reasoning about informative speakers—in that condition. On the other hand, if no reliable difference between model fits is found, this would suggest that listener responses in this task are not driven by the pragmatic reasoning formalized by RSA.

Experiment

In order to investigate the influence of (a) exposure to the task, and (b) experience as a speaker, on the behavior of listeners in language games, the current study uses a 2×2 design (Exposure: First-trial vs All-trials; Block Order: Listener-first vs Speaker-first). More specifically, we compare listener behavior in the first trial of the study (which is qualitatively similar to a one-shot study) to the mean listener behavior across all trials. This Exposure factor is then crossed with Block Order, a between-subjects manipulation wherein half of the participants first play a block of trials in the role of the listener, and half first play a block of trials in the role of the speaker.

Methods

Participants 157 participants were recruited via Amazon's Mechanical Turk and were compensated $0.50. Participants' IP addresses were limited to US addresses only. Only participants with a past work approval rate of at least 90% were accepted. Individuals were not allowed to participate more than once. 24 participants (15.3%) were excluded because they self-identified as non-native or non-fluent English speakers. An additional 13 participants (8.3%) were excluded because they did not meet a 60% accuracy threshold in either the Speaker or Listener task. Inaccuracy in the Speaker task was defined as selecting a word whose literal meaning did not match the target object. Inaccuracy in the Listener task was defined as selecting a referent that did not correspond to the literal meaning of the given word. The remaining 120 participants were randomly assigned to either the *Listener-first* ($N = 60$) or *Speaker-first* ($N = 60$) condition.

Procedure Participants in all conditions were introduced to an interlocutor ("Robert") and were told that they would play a referential communication game that was divided into two blocks. In one block participants played the *Speaker* and in

the other they played the *Listener* (Figure 1). Block order depended on the between-subjects condition (Speaker-first, Listener-first). On each trial participants saw three objects which differed systematically along two dimensions: shape (fish, boot, table, mitt) and color (blue, green, orange, purple). Before the experiment began, participants were shown a display which familiarized them with the shapes, colors, and how they would be labeled. The familiarization display was labeled *Instructions* and said, "These are the objects you will be talking about:", followed by a row of small gray icons depicting the four possible shapes. Then, "And they can appear in these colors:", followed by a row of paint-drop icons depicting the four possible colors.

On Speaker trials, one object (the target) was indicated by an arrow. Speakers were told, "Imagine you are talking to Robert and you want him to pick out the object indicated by the arrow (but Robert can't see the arrow). If you can only use **one word**, which word would you say?" Speakers entered their response into a text input box that only accepted 8 words (fish, boot, table, mitt, blue, green, orange, purple). On Listener trials, the objects were labeled A, B, and C. Listeners were told, "Robert wants you to pick one of the objects above but he can only say one word. He says: [word]. Which object do you think he is talking about?" The category of the given word (shape or color) was counterbalanced across trials. Listeners made their choice by clicking one of three radio buttons labeled A, B, and C. Listeners were further instructed, "Be careful because occasionally Robert's message may become garbled during transmission and you will see "#######" instead of a word. In these cases, just do your best to decide which object you think he is talking about." These instances served as *Salience* trials, used for measuring visual/referential salience. Before the Speaker block, participants completed one practice trial. The Listener block was preceded by two practice trials (one Listener trial, one Salience trial). All practice trials contained three objects which each had a unique color and shape.

Materials Participants saw 42 experimental trials divided into a Speaker block and a Listener block. Speaker blocks contained 18 trials, of which 6 were critical trials and 12 were fillers. On critical Speaker trials (Figure 1), the visual context contained a *pragmatic referent* (e.g., blue boot), a *color competitor* (e.g., blue fish), and a *shape competitor* (e.g., green boot), and the pragmatic referent was indicated as the target. Six of the filler trials used the critical display type (i.e. the objects had the same configuration of features as in the critical trials), however one of the competitors was designated as the target. We refer to these as $filler_{critical\ display}$ trials. The remaining six filler trials contained different configurations of object features (e.g., blue mitt, orange mitt, green fish), but the designated target (e.g., blue mitt) could always be referred to unambiguously (e.g., "blue"). Unique displays were generated for each trial by randomly selecting a target object (e.g., blue boot) and then generating a set of competitors based on the systematic constraints for each condition. For all participants, the first Speaker trial was a critical trial. The remaining Speaker trials were pseudo-randomly mixed and counterbalanced such that: (a) critical trials were always preceded by a $\text{filler}_{\text{critical display}}$, (b) critical trials were evenly distributed across the list, (c) the target appeared an equal number of times in each position (left, middle, right), (d) the target position on trial n was different than the target position on trial $n-1$, and (e) no objects in trial n had the shape or color of the target on trial $n-1$ nor the complementary feature of the competitor from trial $n-1$.[1]

Listener blocks contained 24 trials, of which 6 were critical and 12 were fillers. These context types used the same configuration of object features as in the Speaker trials and participants observed either a color word (e.g., "blue") or a shape word (e.g., "boot"). The remaining six trials in the Listener block were *Salience* trials, three of which used the critical display type and thus served as an empirical measure of the prior likelihood of referring to an object. The remaining three Salience trials used filler display types. Unique displays were generated for each trial using the same constraints as for the Speaker task, plus: (f) an equal number of color and shape words was used in each list, and (g) if trial $n-1$ was a Salience trial then the objects in trial n did not contain any shapes or colors used in trial $n-1$. Importantly, because previous one-shot studies have shown that listeners strongly prefer the shape competitor over the pragmatic referent when given a shape word (Qing & Franke, 2015; Sikos et al., 2021), we constrained the first Listener trial to be a critical shape-word trial.

Behavioral Results

Analyses were conducted on critical trials only. Inaccurate responses (as defined above) were excluded (Listener task: 1.0%; Speaker task: 0.7%). Figure 2 presents the results from the (A) Listener, (B) Salience, and (C) Speaker tasks. Error bars represent binomial 95% confidence intervals (using the 'rstatix::binom_test' function in R on default settings). The dashed lines represent chance.

Listener Task Figure 2A shows the proportion of pragmatic responses by block order and the observed word (shape, color). Responses were considered to be pragmatic if listeners chose the pragmatic referent (e.g., *blue boot*; see Figure 1) over the shape competitor (e.g., *green boot*) when given a shape word (e.g., "boot"), and if listeners chose the pragmatic referent over the color competitor (e.g., *blue fish*) when given a color word (e.g., "blue"). Only the proportion of pragmatic responses is shown because pragmatic and competitor responses sum to 1. Within each panel, the left dot represents the mean proportion of pragmatic responses in the *First* critical Listener trial (shape word: trial 1, color word: trial 6), while the right dot represents the mean proportion of prag-

[1]Due to the many interdependencies among these constraints, a single pseudorandomized trial order was used for all participants. However, the specific shapes and colors for each object were randomly generated for each trial.

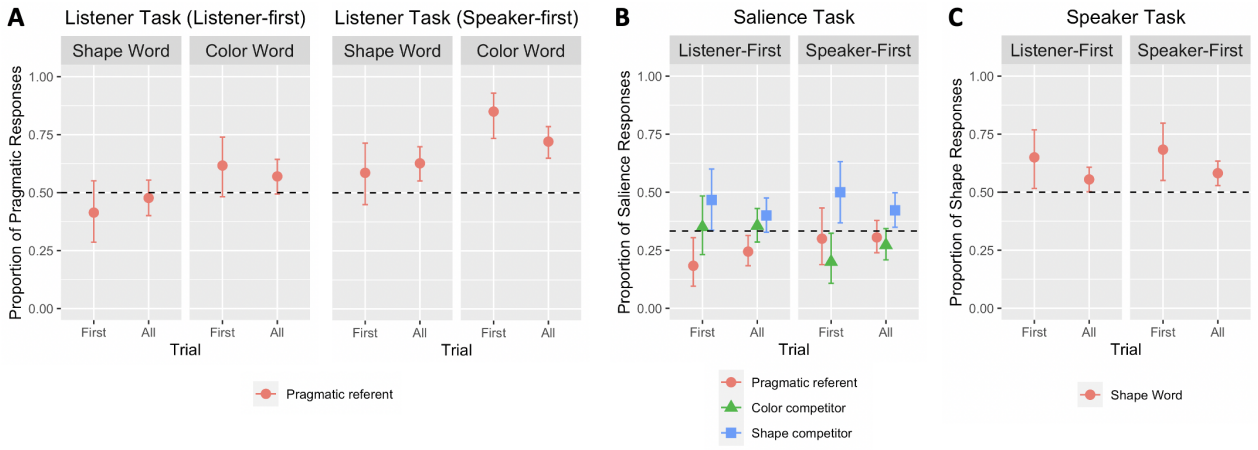

Figure 2: Human Judgments in the Listener task (A), the Salience task (B), and the Speaker task (C). We plot the proportion of responses by block order (Listener-first, Speaker-first) and the observed word (shape, color). Error bars represent binomial 95% confidence intervals and the dashed lines represent chance.

matic responses across *All* critical trials in which listeners observed the given word (shape or color).

Three clear patterns emerge. First, simply increasing the number of trials (Exposure: First-trial vs All-trials) did not lead to a greater proportion of pragmatic responses. Second, listeners had a greater preference for the pragmatic referent in the Speaker-first condition than in the Listener-first condition. Third, listeners had a greater preference for the pragmatic referent in the color-word condition than in the shape-word condition. These observations were confirmed statistically. To assess the effect of the exposure manipulation, we fit a binary logistic regression model using listener responses (pragmatic referent, competitor) as our dependent measure and exposure as a predictor. No effect of exposure was found when collapsing across block order ($p = .60$), nor when looking separately at the Listener-first ($p = .89$) or Speaker-first ($p = .35$) conditions. To assess whether block order and the observed word modulated listeners' initial responses, we fit a binary logistic regression model to the First-trial data using listener responses as our dependent measure, with the observed word (shape, color) and block order (Listener-first, Speaker-first) as predictors. Results revealed main effects of both the observed word ($p < .05$) and block order ($p < .01$). However, the interaction between block order and the observed word was not reliable ($p = .34$). The same analysis on the All-trials data revealed a similar pattern: a marginal main effect of the observed word ($p = .08$), a main effect of block order ($p < .01$), and the interaction was not significant ($p = .86$). A likelihood ratio test determined that adding trial number as a predictor in the All-trial analysis did not improve the model fit, thus we elected to analyze the First-trial data separately.

In addition, the fact that most of the error bars in Figure 2A include 0.5 (chance) suggests that listeners rarely had a reliable preference for the pragmatic referent over the competitor. This observation was confirmed via separate exact binomial tests for each listener-word $\times$ block-order condition. Results for the First-trial data indicate that listeners' initial responses only showed a reliable preference for the pragmatic referent in the color-word, speaker-first condition ($p < .0001$; color-word, listener-first: $p = .09$; shape-word, speaker-first: $p = .24$; shape-word, listener-first: $p = .24$). The same analyses on the All-trial data revealed that listeners reliably preferred the pragmatic referent in both speaker-first conditions (color-word: $p < .0001$; shape-word: $p < .001$), but neither of the listener-first conditions reached significance (color-word: $p = .07$; shape-word: $p = .60$). Taken together, these findings suggest that participants, on average, require some experience with the speaker role before they begin to respond pragmatically as listeners. We speculate that when participants first play the role of the speaker, they gain a better understanding of what it means to be informative in the current task. Consequently, when the Speaker-first participants then play the role of the listener, they are more likely than Listener-first participants to infer that the speaker intended to refer to the pragmatic referent.

Salience Task Figure 2B shows the proportion of responses in the Salience task for the pragmatic referent (e.g., *blue boot*), the color competitor (e.g., *blue fish*), and the shape competitor (e.g., *green boot*). Consistent with previous work (Qing & Franke, 2015), participants had an overall preference for the shape competitor (i.e., the object with a unique color: 0.41; color competitor: 0.31; pragmatic referent: 0.28). To test whether block order and trial number had an effect on Salience responses, we fit a multinomial logistic regression model to the All-trials data using Salience responses (pragmatic referent, color competitor, shape competitor) as the dependent measure, with block order and trial number as the predictors. No reliable effect of block order was found (*ps*

Table 1: Speaker results comparing the critical condition and the filler$_{\text{critical display}}$ condition. Columns represent choice options: color word, shape word. Rows present the distribution of responses when the target was the pragmatic referent (top), color competitor (middle), or shape competitor (bottom).

	Listener-First			Speaker-First		
Target	N	color	shape	N	color	shape
pragm. ref. 	355	0.45	0.55	354	0.42	0.58
color comp. 	149	0.18	0.82	146	0.16	0.84
shape comp. 	207	0.60	0.40	205	0.74	0.26

$>$ .20). However, a significant main effect of trial number was found on the shape competitor / pragmatic referent comparison ($\beta = 0.09$, $p < .01$). This finding suggests that the shape competitor became less salient—and the pragmatic referent more salient—as participants gained experience with the task.

Speaker Task Figure 2C shows the proportion of shape word responses (e.g., "boot") given by speakers in critical trials (i.e. wherein the pragmatic referent was the target) for the Listener-first (left) and Speaker-first (right) conditions. As in previous work (Frank et al., 2016; Frank & Goodman, 2012; Qing & Franke, 2015), speakers had an overall preference for using shape words to describe the pragmatic referent (shape: 0.57, color: 0.43; exact binomial test: $p < .001$), despite the fact that shape and color words are equally informative in the critical contexts. Although this preference decreased numerically over the course of the study, a binary logistic regression model using speaker responses (color word, shape word) as the dependent measure and block order and trial number as a predictor, revealed that neither block order ($p = .52$), nor trial number ($p = .24$), nor their interaction ($p = .76$) were reliable.

To compare speakers' preferences in the critical condition to cases in which one word (e.g., color) was more informative than the other (e.g., shape), we tabulated speaker choices across both the critical condition and the filler$_{\text{critical display}}$ condition (in which the speaker's target was one of the competitor objects; e.g., the *green boot*). Table 1 presents these results for the Listener-first and Speaker-first conditions. The first row shows speaker's preference for using a shape word over a color word in the critical condition, across the Exposure conditions. The bottom two rows show that speakers strongly preferred the more informative word, regardless of whether it was a shape or color word. These results suggest that speakers are highly informative in this task, regardless of whether they see a single trial or multiple trials, or whether they first play the role of the Listener or the Speaker.

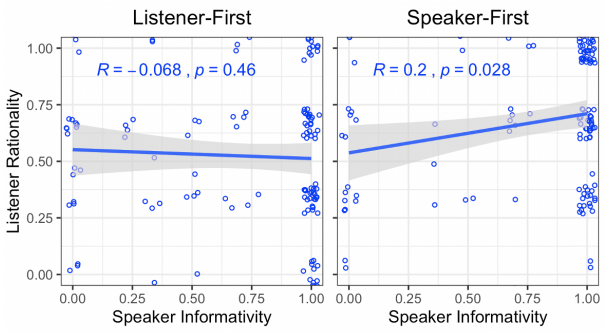

Figure 3: Correlation of Speaker Informativity and Listener Rationality (R: Spearman's rho correlation).

Speaker Informativity vs Listener Rationality As a final behavioral measure of pragmatic reasoning, we assessed whether an individual's tendency to be informative in the Speaker task was correlated with the degree of pragmatic behavior they showed in the Listener task (Figure 3). We computed Speaker Informativity for each participant as the proportion of trials in which they used the more informative word in the filler$_{\text{critical display}}$ condition. Listener Rationality was computed for each participant as the proportion of critical Listener trials in which they chose the pragmatic referent over the competitor. Spearman's rho correlation coefficient was used to assess the relationship between Speaker Informativity and Listener Rationality. No correlation was found for the Listener-first condition ($r_s = -0.07, p = .46, N = 120$). In contrast, a significant correlation was found in the Speaker-first condition ($r_s = 0.20, p < .05, N = 120$). A Fisher z-transformation revealed that the difference between these correlations was significant ($z = 2.09, p < .05$). Although these results are not conclusive, they suggest that *a priori*, there is no relation between how rational individuals are as listeners and how informative they are as speakers, but after engaging in the Speaker task first, the more informative speakers become the more rational listeners.

Model Evaluation

We model our behavioral results within the RSA framework (see Eq. 1). RSA assumes that speakers choose utterances according to their utility, which is defined in terms of word specificity: speakers are more likely to use word w to the extent that it reduces referential uncertainty (Frank & Goodman, 2012). Consistent with the results reported by Frank and Goodman (2012), participant choices in the Speaker task were highly correlated with RSA's predictions for informative speakers in both the Listener-first ($r = 0.99$) and Speaker-first ($r = 0.98$) conditions.[2]

[2]Based on the likelihood function reported by Frank and Goodman (2012) that assumes a single level of recursion depth, a rationality degree of $\alpha = 1$, and a constant cost function (see Frank & Goodman, 2012, Supplemental Materials). The correlation between

Table 2: Model evaluation results from the Listener task comparing RSA to the baseline literal listener model (LL). r: Pearsons's correlation; cocor-p: p-value for comparison of overlapping dependent correlations.)

		Listener-First					**Speaker-First**				
Dataset	**Model**	r	R^2_{adj}	t	p	cocor-p	r	R^2_{adj}	t	p	cocor-p
First-trial	RSA	0.90	0.79	6.59	$< .0001$	$\}$ .43	0.99	0.98	23.07	$< .0001$	$\}$ $< .0001$
	LL	0.94	0.87	8.65	$< .0001$		0.85	0.69	5.08	$< .001$	
All-trials	RSA	0.99	0.97	26.64	$< .0001$	$\}$ $< .0001$	0.98	0.95	21.56	$< .0001$	$\}$ $< .0001$
	LL	0.90	0.79	9.46	$< .0001$		0.80	0.62	6.19	$< .0001$	

We calculate RSA's predictions by combining empirically observed speaker behavior, $P(word|target,C)$ (i.e. the proportion of shape/color word responses to a target in the critical display type C; see Table 1), and empirically observed salience ratings, $P(target)$, in order to predict listener behavior, $P(target|word,C)$:

$$\textbf{RSA: } P(target|word,C) = \frac{P(word|target,C)P(target)}{\sum_{r\in C} P(w|r,C)P(r)} \tag{2}$$

To evaluate the contribution of RSA's pragmatic component (i.e. the speaker model) in predicting listener behavior, we contrast the RSA predictions with those from a baseline literal listener model (LL; see Sikos et al., 2021) that replaces RSA's pragmatic component with a truth-condition function that simply determines whether the literal meaning of the given word fits the referent or not:

$$\textbf{LL: } P(target|word,C) = \frac{[\![word]\!](target)P(target)}{\sum_{r\in C}[\![word]\!](r)P(r)} \tag{3}$$

where $[\![w]\!]$ defines the meaning of word w as a function from the set of all possible referents R to binary truth values ($[\![w]\!] : R \rightarrow \{0,1\}$), such that it returns 1 if the meaning of word w (e.g., "blue") fits the referent (e.g., *blue boot*), and 0 otherwise (e.g., *green boot*). The baseline literal listener model thus predicts that if multiple objects can be described by a given word, listeners will distribute their choices across those objects, weighted by the prior alone.

Table 2 shows the fit of each model's predictions to observed listener responses for the First-trial data (upper two rows) and All-trials data (lower two rows), separated by Block Order. To statistically compare model fits we used the *cocor()* function for comparing overlapping dependent correlations (Diedenhofen & Musch, 2015) with an alpha of 0.05 in the statistical software package R (version 3.6.1; `https://www.R-project.org/`) and we report Hittner et al.'s (Hittner, May, & Silver, 2003) modification of Dunn and Clark's z (Dunn & Clark, 1969).[3]

In the Listener-first condition (left), no reliable difference between model fits was found for the First-trial data. This finding is consistent with the limited evidence for pragmatic reasoning found in previous studies using the one-shot paradigm (Frank et al., 2016; Franke & Degen, 2016; Qing & Franke, 2015; Sikos et al., 2021). For the All-trials data in the Listener-first condition, we do observe a significant difference between the model fits, with RSA outperforming the literal listener model. Interestingly, this finding suggests that RSA's pragmatic component (i.e., the speaker model) contributes to predicting listener behavior, even though the behavioral data did not show an increase in pragmatic responses in the All-trial data versus the First-trial data.

A different pattern of model fits is evident in the Speaker-first condition (Table 2, right). For the First-trial data, we find a reliable advantage for RSA, relative to the First-trial, Listener-first data, which is consistent with the increase in pragmatic behavior found across these conditions in the behavioral results, and a decrease in the utility of salience, as revealed by the diminished performance of the LL model. In the All-trials data, RSA performance stays high, whereas we again observe a decrease in performance of the literal listener model, indicating that the pragmatic component critically contributes to explaining listener behavior. Taken together, these findings provide further evidence that experience with the speaker role leads listeners (on average) to behave more pragmatically.

General Discussion

The results of a number of studies suggest that listeners may not behave as pragmatically as previously assumed in one-shot web-based language games (Frank et al., 2016; Franke & Degen, 2016; Qing & Franke, 2015; Sikos et al., 2021). To investigate whether this finding may be due to the one-shot paradigm itself, we tested whether listeners' pragmatic reasoning is enhanced by greater exposure to the task, and/or experience with being a Speaker in this setting. In addition to replicating previous findings that listeners show limited pragmatic behavior in the one-shot version of the task (when observing a shape word), we find that while increased experience with the (Listener) task improves RSA's performance in predicting listener responses, it does not lead to more pragmatic responses. In contrast, our results indicate that listeners who first experienced the speaker role do have an increased

participant choices and predictions for informative speakers remains high across all values of α between 0 and 10 ($rs > 0.97$).

[3]Many tests have been proposed for comparing overlapping dependent correlations. For a detailed discussion of these competing tests, see Diedenhofen and Musch (2015) and the references therein. The *cocor()* function computes the results of ten different tests. Although we report Hittner et al.'s z, a minimum of eight tests were in agreement for each of the comparisons below.

tendency towards pragmatic reasoning, above and beyond simply increasing the number of trials: behavioral results reveal an increased preference for the pragmatic target in the Speaker-first condition, compared to the Listener-first condition, and model evaluation results confirm that RSA outperforms a baseline Literal Listener model in predicting listener behavior. In addition, we found that the pragmatic referent became more salient as participants encountered more trials, which may explain the overall high correlations of both RSA and LL, in particular in the Listener-first, All-trials condition, as both models incorporate a salience component.

As in previous studies, we observe a clear distinction between shape words and color words: Speakers are overall more likely to use a shape than a color word, and listeners are overall less likely to select the pragmatic referent when observing a shape word. These findings are often attributed to a difference in expectancy between nouns (shape) and adjectives (color) (e.g., Qing & Franke, 2015; Sikos et al., 2021). In contrast to previous studies, however, we find a significant preference for the pragmatic referent after observing a shape word, across all trials in the Speaker-first condition, suggesting that listeners are able to reason pragmatically about observed nouns. One factor that may contribute to this effect is the (non-significant) trend towards a uniform salience distribution over multiple trials, suggesting reduced influence of the shape competitor.

Finally, we observe a correlation between an individual's tendency to be informative in the Speaker task and the degree of pragmatic behavior they show in the Listener task, but only for Speaker-first participants. When considering the Listener-first findings, we find no correlation between pragmatic listener behavior and an individual's subsequent tendency to speak informatively. This result suggests that when participants first engage in the Speaker task, it is the more informative speakers who benefit to a greater extent, rather than resulting in improvements across the board.

In summary, our results confirm the observation put forward by Sikos et al. (2021) that the high correlation between RSA's predictions and listener behavior reported in one-shot experiments (e.g., Frank & Goodman, 2012) is primarily driven by non-pragmatic factors, such as the incorporation of literal meaning and the prior probability of referring to the referent. Importantly, however, our results suggest that the role of RSA's pragmatic component, which reasons about informative speakers, is enhanced not only when listeners have experience with the task, but particularly when they have experience as the speaker. This seems especially relevant for language game settings with which participants are likely unfamiliar. These findings demonstrate that familiarity with the communicative setting can influence the degree of rationality that listeners realize. Correlations between Speaker Informativity and Listener Rationality support this conclusion, showing that participants who first have experience as (informative) speakers are more likely to be pragmatic listeners. In other words: *Good speakers become better listeners*.

References

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*. doi: 10.3765/sp.9.20

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591–621. doi: 10.1037/rev0000186

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PlOS ONE*, *10*(4), e0121945.

Dunn, O. J., & Clark, V. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association*, *64*(325), 366–377.

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) doi: 10.31234/osf.io/f9y6b

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

Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PLOS ONE*, *11*(5), e0154854.

Hittner, J. B., May, K., & Silver, N. C. (2003). A monte carlo evaluation of tests for comparing dependent correlations. *The Journal of general psychology*, *130*(2), 149–168.

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.

Qing, C., & Franke, M. (2015). Variations on a bayesian theme: Comparing bayesian models of referential reasoning. In *Bayesian natural language semantics and pragmatics* (pp. 201–220). Springer.

Scontras, G., & Goodman, N. D. (2017). Resolving uncertainty in plural predication. *Cognition*, *168*, 294–311.

Sikos, L., Venhuizen, N. J., Drenhaus, H., & Crocker, M. W. (2021). Reevaluating pragmatic reasoning in language games. *PLOS ONE*, *16*(3), 1-33. doi: 10.1371/journal.pone.0248388

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.