People do not engage in ad-hoc reasoning about alternative messages when interacting with a literal speaker

Alexandra Mayn[1], **John Duff**[1], **Natalia Bila**[1] and **Vera Demberg**[1,2]
{amayn, jduff, nbila, vera}@lst.uni-saarland.de
[1] Department of Language Science and Technology, Saarland University
[2] Department of Computer Science, Saarland University

Abstract

Derivation of pragmatic inferences typically assumes that both interlocutors behave rationally, as described by the cooperative principle. However, real-world communication often involves speakers who cannot behave fully rationally due to factors such as limited language proficiency, high cognitive load, or insufficient reasoning skills. In such cases, listeners may adjust their inferences to account for the speaker's limitations. In this study, we investigate whether participants engage in ad-hoc reasoning about alternative messages available to the speaker when the speaker is explicitly literal. Our findings reveal that people overwhelmingly do not do so and instead behave as literal listeners, and that nudging participants to consider alternative messages does not improve performance. This suggests that while people readily consider the speaker's intent, they do not tend to engage in ad-hoc reasoning about probabilities of alternative messages in the absence of a rational speaker.

Keywords: pragmatics; partner effects; bounded rationality

Introduction

The cooperative principle, introduced by Paul Grice (1975), states that communicators are expected to adhere to rational communicative principles to enable successful communication. From the listener's perspective, this involves making inferences about the speaker's intent and alternative utterances the speaker could have used to express the intended meaning.

While Gricean pragmatic theory typically models listeners who assume that the speaker is rational, there are many situations in which that may not be the case. Indeed, it has been shown that listeners flexibly adjust their interpretations when they believe that the speaker is not able to behave rationally due to having a limited language proficiency (Ip & Papafragou, 2023), a language impairment (Grodner & Sedivy, 2011), or not very advanced reasoning skills (Mayn et al., 2025).

In a recent study, we found that participants were more likely to interpret an ambiguous message literally when they thought they were interacting with a child, but drew an implicature when they thought they were interacting with an adult (Mayn et al., 2025). We proposed a model of this phenomenon in the Rational Speech Act framework (RSA; Frank and Goodman, 2012). According to this model, the listener has some uncertainty about whether they are dealing with a literal or a pragmatic speaker, and weighs the two speaker models according to their beliefs. However, even when the speaker is assumed to be perfectly literal, the pragmatic listener model always favors a pragmatic interpretation by considering the relative probabilities of alternative messages available to the speaker. This seems to be at odds with what many people do: they appear to successfully reason about the speaker's intent but do not consider the probabilities of alternative messages, thus deriving literal interpretations.

In that study, most participants seemed to perceive the child speaker as lacking advanced pragmatic reasoning skills, though substantial individual variation suggests that people hold different beliefs about children's communicative abilities. In the current study, we investigate people's reasoning about a speaker who is explicitly presented as literal, a simple computer program which randomly chooses any true message to refer to an object. This design allows us to separate participants' reasoning about alternative utterances from any assumptions about the speaker's intent or rationality. We ask whether participants will successfully incorporate the information about alternative messages available to the speaker into their inference, leading them to favor a pragmatic interpretation, or whether they will derive a literal intepretation instead. A pragmatic interpretation, in this context, requires adjusting the inferred likelihood of each referent based on how many literally true messages apply to it, corresponding to $P(object|utterance) \propto P(utterance|object)$, whereas a literal interpretation corresponds to assigning equal probability to all literally true options.

Some evidence that people might struggle to correctly incorporate probabilities of alternative messages into their interpretation comes from literature on Bayesian reasoning errors in other domains. It has been shown that people often fail to apply Bayesian principles when solving probability problems and instead rely on heuristics (Tversky & Kahneman, 1993) or combine different pieces of information, such as base rates and false alarm rates, linearly, instead of applying Bayes' rule (Stengård et al., 2022). Fox and Levav (2004) showed that their participants overwhelmingly did not solve Bayesian reasoning problems correctly due to incorrectly partitioning the probability space, but nudging them to consider all possible events significantly increased Bayesian reasoning.

We find that, when presented with an explicitly literal speaker, people overwhelmingly do not perform ad-hoc reasoning about alternative messages and instead provide responses consistent with a literal interpretation. Furthermore, nudging people to consider all possible messages does not

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

result in more pragmatic responding. Our findings suggest that while reasoning about speaker's intent is natural, reasoning about probabilities of alternative messages, at least in the absence of speaker's intent, is effortful and is not default behavior.

We discuss the fact that the RSA framework currently does not allow for decoupling reasoning about alternative messages from reasoning about the speaker's intent or rationality and that it may be needed to model a situation where the speaker's intent is successfully considered but the probabilities of alternatives are not.

Background: Reference game

The task participants completed is a reference game, a simple signaling game frequently used to test the predictions of RSA models. Participants were told that they would play a communication game with a simple computer program called *basic_message_picker*, which randomly selects any true message to refer to an object. Participants briefly practiced selecting messages as if they were the computer program and received feedback to ensure they understood its expected behavior.

On each trial, participants saw three objects, a pictorial message that *basic_message_picker* purportedly selected to refer to one of them, and a set of pictorial messages that were available to it. The set of available messages remained the same for the whole experiment. Participants then indicated the likelihood that each of the three objects was the intended referent by distributing 100 points between the objects using sliders.

On critical trials, as shown in Figure 1, the message is true of two objects. However, for one of the objects (the target), one of its features is inexpressible (there is no message for "square"), whereas the other object (the competitor) could also be referred to with another message ("triangle").

In the case when the speaker is expected to be rational (e.g., another adult participant), it has been repeatedly shown that the majority of participants prefer the target and reason that if the speaker had wanted to refer to the competitor, they could have used the more optimal message, "triangle", which would have been unambiguous (Frank & Goodman, 2012; Franke & Degen, 2016; Mayn et al., 2025). This reasoning is formulated by the pragmatic listener RSA model L_2, who reasons about the pragmatic speaker S_1, who, in turn, reasons about the literal speaker L_0.

In this case, however, the speaker is explicitly presented as literal. Nevertheless, it is still possible to make a pragmatic inference and prefer the target by reasoning not about the speaker's intent but about the distribution of alternative messages available to the speaker. There is only one way to refer to the target (in the example in Figure 1, it is the message "green"), and two ways to refer to the competitor ("green" and "triangle"). Assuming that the speaker is equally likely to refer to either object, the message "green" is then twice as likely to be referring to the target than to the competitor, resulting in the target probability of $\frac{2}{3}$. This is predicted by all pragmatic listener models more complex than L_0.

It is also possible that a listener does not consider the distribution of alternative messages and interprets the message literally instead, resulting in assigning an equal probability (50) to the target and to the competitor. This is what the literal listener model L_0 would predict.

Finally, it is possible that a listener tries to perform pragmatic reasoning but makes an error in probabilistic computations about alternative messages. There is a lot of evidence from other domains that people struggle with performing Bayesian computations, therefore, it could also be the case in this setting. This would also result in assigning equal ratings to the target and to the competitor (50%) but it is a distinct process from L_0 in that pragmatic reasoning is attempted but it involves a probabilistic computation failure, whereas L_0 models a case where no pragmatic reasoning is attempted.

There are two sources of information one could consider when deriving an inference: the speaker's intent or rationality and the available alternative messages and their probabilities. A literal listener model L_0 considers neither, and a pragmatic listener model L_2 considers both. Importantly, any recursive listener RSA model which has an internal speaker model (i.e., a model more complex than L_0) will automatically correctly incorporate reasoning about alternative messages into the inference. Thus, modeling a listener who successfully reasons about the speaker's intent but does not about alternative messages presents a challenge for the RSA.

Importantly, we do not claim that interacting with a literal (or otherwise not very rational) speaker is a case RSA should necessarily be equipped to model. It is, after all, a *Rational* Speech Act model, which formalizes the cooperative principle, at the core of which lies the idea that both interlocutors are being cooperative, and it models an idealized scenario of near-perfect rationality of both interlocutors. Still, we consider the question of how rational, or near-rational, listeners accommodate not-so-rational speakers to be relevant for a realistic model of pragmatic reasoning, since there are many situations where speakers may not be in a position to behave rationally given limited language proficiency, high cognitive load or not sufficiently developed reasoning ability. Because this is such a commonplace setting, it may be one which we may want to model, whether in the RSA model or in another framework. Currently, in RSA there is no way to derive the prediction that people may reason about the speaker but not correctly incorporate alternative messages into their reasoning. We return to this point in the discussion.

Experiment

The experiment investigated participants' interpretations of messages uttered by a literal speaker.[1] The experiment was hosted on LingoTurk, an open-source crowdsourcing server system for psycholinguistics experiments (Pusse et al., 2016).

[1]The preregistration for this study is available at https://osf.io/dh8wm.

Participants

94 native English speakers with an approval rate of 95% and above, who were recruited on the crowdsourcing platform Prolific, completed the experiment.

Procedure

The experiment consisted of three phases: Block 1, training, and Block 2. Participants were randomly assigned to one of two conditions, No training or Training. The order of trials was the same for all participants.

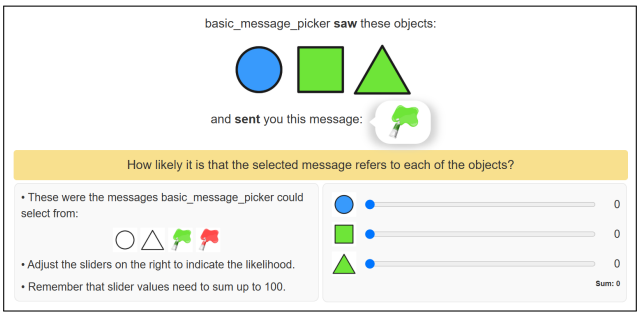

Figure 1: Example critical trial (Block 1)

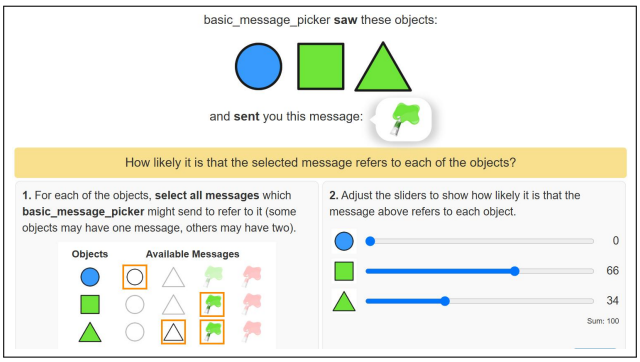

Figure 2: Example critical trial in the Training condition (training and Block 2)

Block 1. All participants completed 8 critical trials, 12 unambiguous (where the message clearly identified a single referent) and 4 ambiguous (where the target and competitor were identical, making it impossible to distinguish between them based on the message) filler trials, interleaved. Because Block 1 was practically identical across groups[2], performance on the 8 critical trials in this block serves as a baseline to evaluate the effect of the subsequent manipulations. At the end of Block 1, participants saw one critical trial again and were prompted to briefly explain, in a textbox, why they decided to put the sliders in those positions.

Training block. After Block 1, participants in the Training condition completed 4 additional reference game trials containing a secondary task, intended to increase awareness of alternative messages which could be used to refer to each object. Before assigning probabilities using sliders, participants had to complete an additional step. For each object, they were presented with a list of all the messages that were available to *basic_message_picker*. Their task was to select all messages that were true for each object (Figure 2). Participants received immediate feedback indicating whether their selection was correct, and they were unable to proceed to the probability sliders unless they had correctly identified all true messages. They only received feedback on the message selection and not on the sliders. If some participants provide low target ratings on critical trials simply because of inattention to the asymmetrical distribution of available messages, this extra step should support improved performance by drawing attention to that distribution. In the No training condition, participants completed an unrelated task (a 10-question version of Raven's progressive matrices) in place of training.

Block 2. All participants then completed a further sequence of 8 critical trials and 16 filler trials, interleaved. In this block, apart from 10 ambiguous and unambiguous filler trials similar to those in Block 1, participants saw 6 filler trials which were more complex. In these trials, the probabilities of the three objects when considering the distribution of available messages should be $^2/_5 : ^2/_5 : ^1/_5$ (2 trials), $^1/_3 : ^1/_3 : ^1/_3$ (2 trials), and $^1/_2 : ^1/_4 : ^1/_4$ (2 trials). These trials were added so that, if there is a positive effect of training, we could investigate whether the understanding generalizes to other situations.

In the No training condition, the procedure in this block was identical to that in Block 1. In the Training condition, participants again had to complete the additional step of choosing all possible messages for every object, as introduced in the training block. However, in this block, they did not receive any feedback on their selections.

Annotation of strategies

Participants' open-ended explanations of their response after Block 1 were annotated following an annotation scheme based on Mayn et al. (2025).

Explanations reflecting reasoning about alternative messages available to the speaker and their respective probabilities were assigned the tag *correct_reasoning*. An example of an explanation from this category is "If the [target] shape is the green square, *basic_message_picker* can only send green, whereas if the shape is the triangle, *basic_message_picker* would send the messages green and triangle with probability 0.5 each. It is therefore more likely that green corresponds to the square."

Explanations indicating belief that the two fitting objects

[2]There was a slight layout difference between the No training and Training conditions. In No training, the available messages were shown below the three possible referents, whereas in Training, they were positioned on the left-hand side of the screen. This change was implemented after an initial test of the Training condition, which used the same layout as No Training. We observed a drop in participants' performance during Block 2, suggesting the training might have been confusing. To address this, we adjusted the layout to reduce clutter on the screen during training and Block 2. The layout changes were small, and the fact that performance in Block 1 is very similar across both conditions suggests that the two conditions remained comparable. Images of the two layouts can be found at https://osf.io/erbn3.

were equally probable were labeled *guess*.

A few participants' explanations revealed that they expected *basic_message_picker* to behave rationally, despite the instructions explicitly stating that it is literal. These were labeled *ascribe_rationality*. An example from this category is "It doesn't make much logical sense to choose green for [referring to] the green triangle when there are two green shapes and you could have chosen the triangle symbol instead."

Responses which suggested that the participant changed their mind upon reflection were labeled *changed_mind*, indicating that their provided explanation did not match the strategy they used in the moment. Explanations which were hard to interpret or which did not reveal anything about the participant's strategy were labeled *unclear*. Participants with these labels were retained for the main analysis.

One participant's reported strategy indicated that they had misunderstood the instructions of the experiment. Their explanation was assigned the tag *misunderstood_instructions* and the participant was excluded from all analyses.

Other tags used in Mayn et al. (2025) are not described here as they did not occur in our data.

Results

12 participants were excluded for accuracy below 80% in unambiguous filler trials in either block, as this seems to suggest a general lack of attention. One further participant was excluded because their reported strategy suggested that they had misunderstood the instructions. Those two exclusion criteria were preregistered. Two participants were excluded because they appeared to have had technical issues: their answers were saved on the server multiple times and were slightly different every time. The remaining 79 participants (41 in the No training condition, 38 in Training) entered the analysis.

First, we classified participants' performance in each block based on the responses on critical trials. We computed the likelihood of participants' responses coming from normal distributions centered around 50, corresponding to a literal interpretation, 66 (or $^2/_3$), corresponding to a pragmatic interpretation, and 100, corresponding to ascribing rationality to the speaker, with *sd*=2. Participants whose likelihood for all three classes was at or below the threshold set based on a pilot study (10^{-30}) were classified as "other". This threshold controls how conservative our classification approach is, i.e., how much deviation from the above-mentioned means we can still confidently consider to belong to that class. We chose a fairly conservative threshold by visually inspecting the data.

Figure 3 shows participants' mean target ratings in each block with their assigned class and annotation of their reported strategy for Block 1. Due to implementation error, we did not obtain participants' reported strategies after Block 2. Analysis of annotations is only supplementary to the main analysis and is meant to provide more support for ratings aligning with people's actual strategies.

In both blocks in both conditions, most participants (65% or more) were assigned to the "50" class, indicating literal responding. We see that very few people were assigned to the "66" class, corresponding to correct reasoning about probabilities of alternative messages: in Block 1, only one person in the No training condition and none in the Training condition, and in Block 2, two people in the No training condition and three people in the Training condition.

	Estimate	SE	t	p
Intercept	16.92	0.81	20.87	$<$ **0.001**
Trial Number	0.06	0.10	0.59	0.55
Block (2 vs. 1)	0.66	0.53	1.26	0.21
Condition (Tr. vs. No tr.)	0.24	1.13	0.22	0.83
Target Position (Left vs. Ctr.)	–0.29	0.49	–0.58	0.56
Target Position (Right vs. Ctr.)	0.15	0.49	0.31	0.76
Message Type (Color vs. Shape)	–0.11	0.22	–0.49	0.63
Block × Condition	–1.28	0.79	–1.63	0.10

Table 1: Regression output.

It appears that people overwhelmingly interpreted ambiguous messages sent by *basic_message_picker* literally, consistent with predictions of an L_0 model, and that nudging participants to consider alternative messages in training did not improve performance. Participants in the Training condition did not struggle with the added message selection step, as evidenced by their high accuracy on message selection in Block 2 (mean proportion correct = 0.95, *sd* = 0.07). This suggests that the lack of improvement is not caused by confusion about the added message selection step, or a failure to accurately enumerate alternative messages, but crucially by not exploiting those alternative messages in the main task.

To verify this observation and gain more insight into the effect of training, we fit a linear mixed-effects regression model to the data. We only used critical trials. The dependent variable was the absolute distance of the rating given to the target from the "correct" answer of $\frac{2}{3}$, computed as the minimum absolute distance from 66 and 67. The dependent variable was regressed onto block (1 or 2, dummy-coded with 1 as reference), condition (no training or training, dummy-coded with no training as reference), trial number (mean-centered), position of the target (left, middle, or right, dummy-coded with middle as reference), message type (color or shape, sum-coded with shape as reference), and the interaction between block and condition. The random effect structure consisted of random intercepts per participant and per item.

For this analysis, we removed the four people in the Training condition marked with crosses in Figure 3 because their mean accuracy was below 40 and their by-trial performance showed inconsistent responding likely due to not paying attention or misunderstanding the task.

The results are reported in Table 1. There is no effect of condition (β = 0.24 (0.13), p = 0.83) or of block (β = 0.66

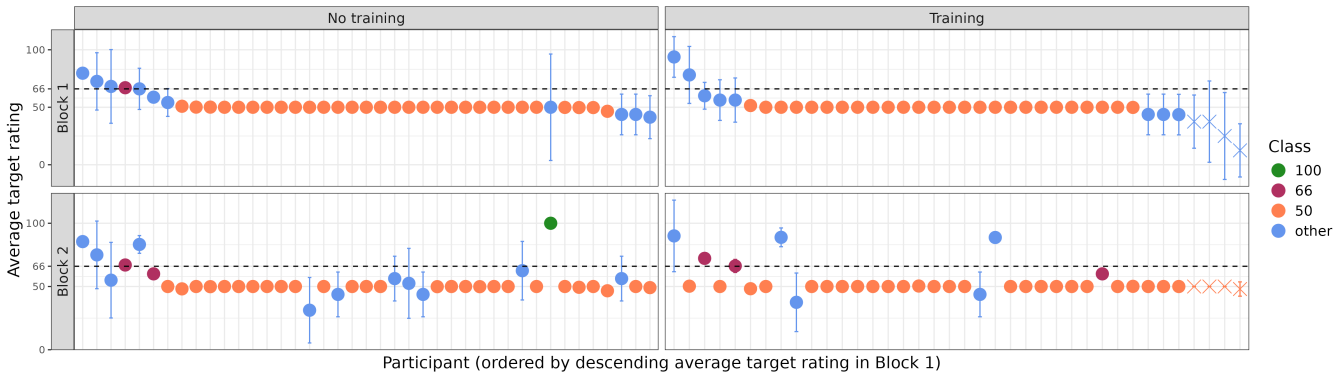

Figure 3: Individual participants' average target ratings on critical trials across blocks in the two conditions, ordered by their average rating in Block 1 (in descending order), with their assigned class.

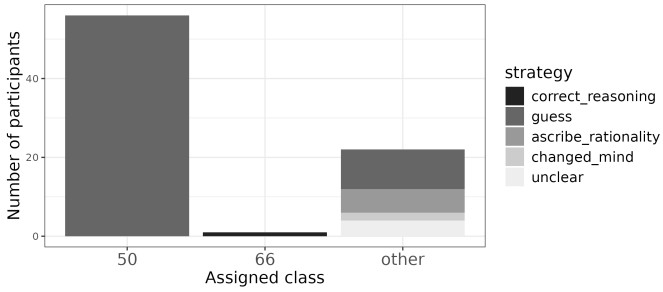

Figure 4: Alignment between the class assigned based on performance and participants' reported strategies for Block 1.

(0.53), $p = 0.21$), suggesting that people were not more likely to derive a correct interpretation with greater exposure. Importantly, the interaction between block and condition is not significant ($\beta = -1.28$ (0.79), $p = 0.10$): participants in the Training condition did not improve more than in the No training condition, meaning that the training did not make participants more likely to correctly weigh alternative messages.

As we mentioned above, for the regression analysis we excluded the four people with suspiciously low target ratings (below 40). If we keep those people in[3], the effect of condition and the interaction of block and condition become significant, suggesting that people in the Training condition were initially further away from the correct rating of $^2/_3$, but got closer after training. However, closer inspection of those four people's performance (Figure 3) reveals that this interpretation is misleading: these participants gave low ratings in Block 1, then consistently rated the target at 50 in Block 2, indicating a shift to a literal interpretation (50), not a pragmatic one (66). Therefore, making the conclusion that the training was helpful is not warranted. We interpret this as evidence that training improved task understanding but not the tendency to derive the pragmatic interpretation.

Next, we take a look at the changes in individual participants' class assignment based on performance between blocks in the two conditions, shown graphically in Figure 5. We see that, in both conditions, there is a large degree of consistency between the assigned classes, suggesting that most participants do not figure out the correct interpretation during the task. In both conditions, the majority of participants fall into the "50" class, and in the Training condition, the main change is that most participants previously classified as "other" join the "50" class, becoming more consistently literal responders. This supports the interpretation that while training helped clarify possible confusion about the task, it mostly led to strengthening of literal responding as opposed to an increase of pragmatic responding.

Finally, we examine the alignment of participants' reported strategies for Block 1 with the class assigned based on their ratings, to assess how closely participants' explanations match their behavior. Figure 4 shows the distribution of reported strategies by class. Since Block 1 is identical across conditions, we pool the data. We see that participants' explanations are generally consistent with class assignment based on ratings, but that a small part of guessers, as well as all participants who expected *basic_message_picker* to behave rationally (*ascribe_rationality*) were classified as "other". The reason that participants whose responses were labeled *ascribe_rationality* were assigned to the "other" and not to the "100" class seems to be that people provided lower ratings than 100, reflecting uncertainty about whether *basic_message_picker* would select the optimal message. This aligns with the observation in Mayn et al. (2025) that even people whose strategy was *correct_reasoning* had some uncertainty about their interlocutors' ability to reason rationally. Overall, this suggests that the classification based on ratings is a fairly reliable proxy for strategy.

Discussion

In this study, we investigated whether people would correctly perform ad-hoc reasoning about alternative messages available to the speaker when the speaker was explicitly literal, a simple computer program which is equally likely to select any true message to refer to an object. We found that participants overwhelmingly did not incorporate probabilities of

[3]This exclusion criterion was not preregistered; we did not anticipate participants giving ratings below 50.

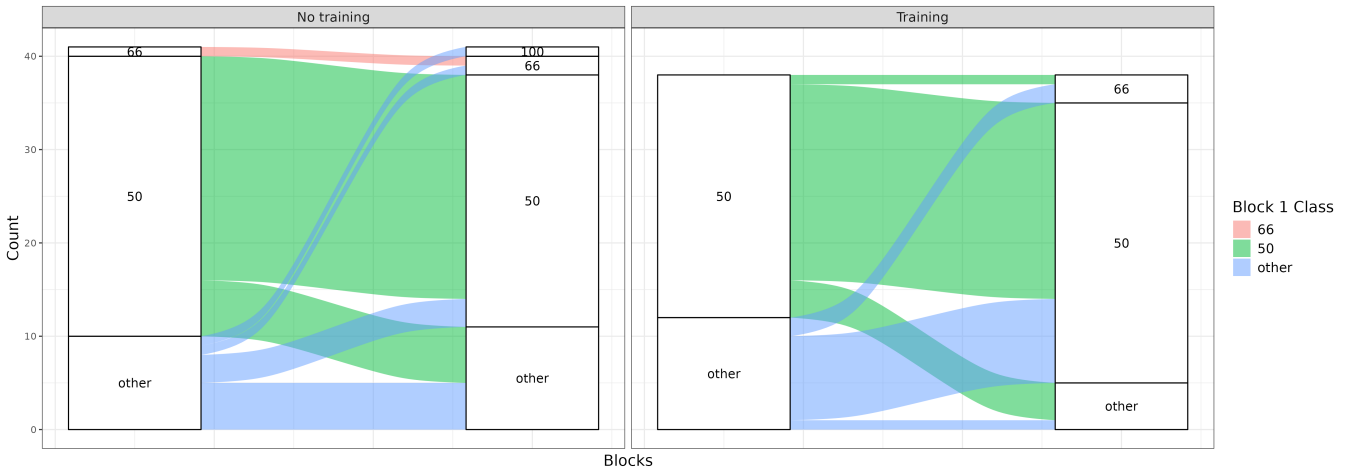

Figure 5: Changes in participants' class assignments based on performance across the two blocks in both conditions.

alternatives into their inferences and instead provided a fully literal interpretation, consistent with predictions of a literal listener RSA model. These findings are in line with those of an earlier study, where we found that when interpreting messages purportedly sent by a child speaker, participants tended to behave as literal comprehenders (Mayn et al., 2025). This is noteworthy, as prior work in the same paradigm has repeatedly shown that with a rational speaker, most participants behaved as pragmatic comprehenders (Degen, Franke, et al., 2012; Franke & Degen, 2016; Mayn & Demberg, 2023).

Interestingly, nudging people to attend to alternative messages through training did not increase the rate of pragmatic interpretations. This suggests that weighing probabilities of messages ad-hoc is a difficult task which is not caused by inattention. This interpretation aligns with literature on errors in Bayesian reasoning in other domains, which shows that people often make errors in probabilistic computations (Fox & Levav, 2004; Starns et al., 2019).

Of course, questions remain about the extent to which these findings transfer to real-life language use. It could be that people do not engage in this kind of reasoning about alternative messages ad-hoc, especially in an unfamiliar scenario, but that they are more successful with familiar linguistic alternatives. It is also possible that when dealing with a not-so-rational speaker, people choose to expend less effort and reason less deeply and fall back on a literal interpretation.

We also observed that some participants expected the speaker to behave rationally and ascribed intention to it despite having been explicitly told that the program is fully literal, again supporting the observation that reasoning about a speaker's intent comes very naturally and sometimes may be the default that needs to be suppressed.

The assumption behind assigning a $\frac{2}{3}$ probability to the target under a pragmatic interpretation is that each of the three objects is equally likely to be referred to. In our design, participants were told that the *basic_message_picker* was given a specific object to refer to on each trial, but they were not told how these targets were selected. As a reviewer pointed out, participants might instead assume that the likelihood of an object being referred to is proportional to the number of expressions available to describe it: the competitor, which can be referred to in two ways, might be assumed to be twice as likely to be referred to as the target, which can only be described with one message. Under this alternative assumption, a target rating of 50 would align with a pragmatic rather than a literal interpretation. While this is a plausible interpretation, we consider it unlikely that most participants adopted such a complex prior. Future work should test whether explicitly stating that all three objects are equally likely to be referred to reduces the proportion of target ratings of 50.

Finally, since it appears that participants generally do not take probabilities of alternatives into account when performing ad-hoc reasoning about a literal speaker, it may be important to develop a cognitive model of reasoning that captures this tendency. While the Rational Speech Act model is primarily a model of pragmatic competence (though there have been some efforts of extending it to account for performance limitations, see e.g. Hawkins et al., 2021) and describes what humans are in theory able to do, it is also important to be able to accurately model performance, especially in cases where it seems to diverge significantly from the predicted competence.

As discussed in Background, in the Rational Speech Act model, any pragmatic listener model more complex than a literal listener will correctly weigh the probabilities of alternative messages, as well as reason about the speaker's intent, whereas a literal listener model will consider neither the speaker's intent nor the probabilities of alternatives. In practice, it could be that these two factors may need to be decoupled in a model, as we see that reasoning about intent is something that comes very naturally to people, whereas reasoning about alternative signals seems to be more effortful. Therefore, it is possible that people are trying to account for alternative messages but failing to do so correctly. Developing a model which captures reasoning about rationality and but does not correctly weigh the probabilities of available alternatives is a potential direction for future work.

Acknowledgments

The authors thank the anonymous reviewers for their helpful comments. They also thank Michael Franke and the audience of XPrag 2024 for helpful discussions. This project is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878).

References

Degen, J., Franke, M., et al. (2012). Optimal reasoning about referential expressions. *Proceedings of SemDIAL*, 2–11.

Fox, C. R., & Levav, J. (2004). Partition-edit-count: Naive extentional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General*, *133*(4), 626–642.

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

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

Grice, H. (1975). Logic and conversation. *Syntax and semantics*, *3*.

Grodner, D., & Sedivy, J. (2011). The effect of speaker-specific information on pragmatic inferences. In E. Gibson & N. J. Pearlmutter (Eds.), *The processing and acquisition of reference*. MIT Press.

Hawkins, R. D., Gweon, H., & Goodman, N. D. (2021). The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive science*, *45*(3), e12926.

Ip, M. H. K., & Papafragou, A. (2023). The pragmatics of foreign accents: The social costs and benefits of being a non-native speaker. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(9), 1505.

Mayn, A., & Demberg, V. (2023). High performance on a pragmatic task may not be the result of successful reasoning: On the importance of eliciting participants' reasoning strategies. *Open Mind*, *7*, 156–178.

Mayn, A., Loy, J. E., & Demberg, V. (2025). Beliefs about the speaker's reasoning ability influence pragmatic interpretation: Children and adults as speakers. *Open Mind*, *9*, 89–120.

Pusse, F., Sayeed, A., & Demberg, V. (2016). Lingoturk: Managing crowdsourced tasks for psycholinguistics. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 57–61.

Starns, J. J., Cohen, A. L., Bosco, C., & Hirst, J. (2019). A visualization technique for Bayesian reasoning. *Applied Cognitive Psychology*, *33*, 234–251.

Stengård, E., Juslin, P., Hahn, U., & Van den Berg, R. (2022). On the generality and cognitive basis of base-rate neglect. *Cognition*, *226*, 105160.

Tversky, A., & Kahneman, D. (1993). Probabilistic reasoning. *Readings in philosophy and cognitive science*, 43–68.