Individual Differences in a Pragmatic Reference Game

Alexandra Mayn (amayn@lst.uni-saarland.de)
Department of Language Science and Technology, Saarland University
Campus C7.2, 66123 Saarbrücken, Germany

Vera Demberg (vera@coli.uni-saarland.de)
Department of Language Science and Technology;
Department of Computer Science, Saarland University
Campus C7.2, 66123 Saarbrücken, Germany

Abstract

While population-level models often provide a good fit to the data, they may mask meaningful individual differences. Exploring individual differences can also be beneficial for gaining a better understanding of the processes that underlie pragmatic phenomena. In this study, we investigate whether the substantial differences in performance on a pragmatic reference game can be traced back to cognitive or socio-pragmatic traits. We observe a significant effect of the ability to inhibit an intuitive response and of abstract reasoning ability. In contrast, we do not find evidence that socio-pragmatic abilities or working memory capacity influence pragmatic responding.

Keywords: experimental pragmatics; individual differences

Introduction

People have been found to differ in the rate at which they draw pragmatic inferences (Heyman & Schaeken, 2015; Yang, Minai, & Fiorentino, 2018; S. C. Fairchild, 2018). In formal probabilistic models of pragmatics (such as Rational Speech Act (RSA) (Frank & Goodman, 2012) and Iterated Best Response (IBR) (Franke et al., 2009)) these differences can be captured via varying the rationality parameter or by manipulating the depth of recursive reasoning that the speaker and the listener perform.

Franke and Degen (2016) argue that the difference originates from depth of reasoning. They conducted an experiment where participants were asked to reason about the potentially ambiguous messages sent by the previous participant, and showed that participants naturally fell into groups whose performance lined up with the predictions of IBR models for listeners of three different reasoning depths. They argued for the usefulness of individual-level modeling in addition to population level modeling.

While Franke and Degen (2016) showed that the individual-level model was able to better account for the data, their model is agnostic about the factors which underlie these differences in pragmatic performance. In this study, we test whether the observed differences in reasoning sophistication can be explained by cognitive or personality factors. If they can, that would provide evidence for the existence of different pragmatic interpreters and shed light on the algorithmic-level processes that are involved in said reasoning. In that case, it would be worthwhile to investigate how stable these reasoning types are over time and to what extent they are specific to this task. If performance cannot be explained by individual differences measures, perhaps the difference between reasoning types emerges from different task strategies or other factors pertinent to the specific occasion when participants completed the task as opposed to stable cognitive or personality differences.

Previous work has explored cognitive and personality-oriented individual differences and their ability to predict pragmatic responding. Yang et al. (2018) observed that people differ in how sensitive they are to the context when deriving scalar implicatures. They found that participants with higher working memory capacity and higher socio-pragmatic abilities were more sensitive to the context. S. Fairchild and Papafragou (2021) investigated the role of executive function and Theory-of-Mind on three pragmatic tasks – scalar implicature, indirect requests, and metaphor, and found that Theory-of-Mind was a significant predictor, whereas working memory was no longer significant once Theory-of-Mind was included. Other studies have found that pragmatic responding can be modulated by nonverbal intelligence (Huettig & Janse, 2016) and attentional control (McVay & Kane, 2012).

In two experiments, we replicate Franke and Degen (2016)'s finding that participants form groups predicted by formally defined reasoning types, and investigate which measures predict pragmatic responding on this task. We find effects of nonverbal intelligence and cognitive reflection ability, but not of working memory or socio-pragmatic abilities.

Background

The task

The main task is an exact replication of Experiment 1 of Franke and Degen (2016). On each trial, participants saw three objects on the display. The objects differed along two dimensions – creatures (green monster, purple monster, and robot) and accessory (red hat, blue hat, and scarf). Participants also saw a message that they were told had been sent by the previous participant. The possible messages were the two monsters and the two hats – robot and scarf are the inexpressible features. The possible messages were always displayed at the bottom of the screen.

In the beginning of the task, participants completed 4 practice trials which employed the speaker's perspective: participants had to select a message to refer to one of three objects on the screen, which was highlighted. Then the actual task

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

began, which consisted of 66 trials, of which 24 were critical and 42 were fillers, distributed as follows:

Half of the critical trials were simple implicature trials. On those trials, the target contained the sampled message and the inexpressible feature along the other dimension. For instance, if the sampled message was the red hat, the target would be the robot with a red hat. The competitor was composed of the message and an expressible feature along the other dimension. Continuing with our example, the competitor could be a purple monster with a red hat. Finally, the distractor was composed of any two features not present in the target or competitor. In our example, it could be a green monster with a scarf (Figure 1, top panel). The simple implicature is predicted to be simple because only one step of reasoning is required to solve it: that the target (robot in our example) cannot be referred to in any other way since *robot* is not an available message.

On complex implicature trials, the target contained the message along with an expressible feature along the other dimension. So if the sampled message was the red hat, the target could be the green monster with a red hat. The competitor contained the message and the other expressible feature along the other dimension. In this example, the competitor would be the purple monster with the red hat. The distractor combined the other target feature not denoted by the sampled message with the inexpressible feature along the other dimension. In our example, the distractor would be the green monster with a scarf (Figure 1, bottom panel). On complex trials, both the target and the distractor can be referred to with two messages, so an additional reasoning step is needed to determine that if the speaker had meant the competitor, she could have used the unambiguous message *purple monster*.

The experimental results of Franke and Degen (2016) (as well as our replication reported here) show that indeed the complex condition is more difficult than the simple one.

Of the 42 filler trials in the study, 24 were identical to the critical trials, except the target was the competitor (6) or distractor (6) from the simple condition, or the competitor from the complex condition (12), unambiguous given the message. For example, there would be a trial like that on top of Figure 1, except the target would be the purple monster with the red hat, and purple monster would also be the unambiguous message. Of the remaining 18 trials, 9 were completely unambiguous (contained each creature and each accessory only once) and 9 were completely ambiguous (contained two identical creatures). Ambiguous trials were included as the random baseline, and unambiguous fillers served for determining exclusion.

Reasoning types

An idea central to Franke and Degen (2016)'s study is that of different pragmatic reasoning types – formal models which describe speakers and listeners and which vary in the depth and complexity of the performed reasoning. Each type of level $n+1$ operates on the assumption that their interlocutor is an n-level reasoner.

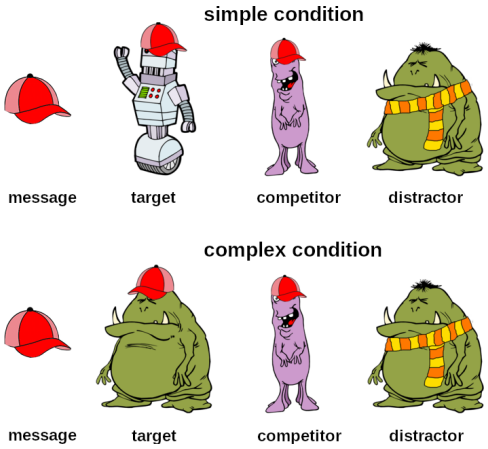


Figure 1: An example of a simple and a complex implicature trial.

The *literal listener* L_0 interprets all messages literally and considers every interpretation that is literally true given the message equally likely. A level-2 listener L_2 is a *Gricean listener* who assumes that their interlocutor is a Gricean speaker S_1. There's also a listener model in between these two types – L_1, who Franke and Degen (2016) call an *exhaustive listener*. These three listener models make different predictions for the reference game. The literal listener L_0 would not be able to solve either the simple or the complex implicature and would assign equal probability to all objects on the display which contain the message. So, continuing with our example in Figure 1, in the simple trial, both the robot (target) and the purple monster (competitor) are wearing the red hat, and L_0 would be at chance in choosing between those. The exhaustive listener L_1 correctly reasons that an unbiased literal speaker S_0 is more likely to refer to the robot since the competitor can be referred to with another message, purple monster. In contrast, neither L_0 nor L_1 will be able to solve the complex implicature trials because both the target and the competitor can be referred to with two different messages. L_2, however, would reason that a pragmatic speaker S_1 would have used the unambiguous message *purple monster* if she had wanted to refer to the purple monster and will correctly identify the target referent.

Cognitive tasks

This section describes the cognitive tasks that we used in the experiments and the rationale for including them.

Need for Cognition Questionnaire (NfC) The Need for Cognition (NfC) questionnaire (Cohen, Stotland, & Wolfe, 1955) assesses how much participants enjoy solving challenging problems. Our intuition was that perhaps more advanced reasoning types would select higher NfC values.

Operation Span (Ospan) There are conflicting findings on the effect of working memory capacity (WMC) on implicature derivation in the literature. S. Fairchild and Papafragou

(2021) found that the effect of WMC on scalar implicature derivation disappeared when a measure of Theory-of-Mind reasoning was taken into account, suggesting that working memory may be engaged for holding the components needed for ToM-reasoning in memory. Yang et al. (2018) report a positive relationship between WMC and sensitivity to context when deriving implicatures. We hypothesized that in case of the reference game, too, the reasoning required for solving the implicatures in the reference game might require sufficient working memory.

We used the automated online version of operation span (Ospan), developed by Scholman, Demberg, and Sanders (2020), where participants verified math equations while holding letter sequences in memory.

Autism Spectrum Quotient Questionnaire (AQ) Yang et al. (2018) found an effect of socio-pragmatic abilities, as measured by the AQ (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001), on context sensitivity when deriving implicatures. We hypothesized that here, too, socio-pragmatic abilities might play a role since people who are higher in autism might be less likely to put themselves in the interlocutor's position and think about why the speaker said what they did, an ability relevant to resolving ambiguity in the reference game.

Raven's Progressive Matrices Test (IQ) Raven's Progressive Matrices was included as a test of nonverbal intelligence. We hypothesized that abstract reasoning ability measured by this task may play a role in deriving implicatures in the reference game. Since the score on as few as 9 items has been shown to correlate almost perfectly with a full-length IQ test (Bilker et al. (2012)), we used a shortened version of the full Progressive Matrices Test consisting of 10 questions of increasing difficulty.

Stroop It could be the case that the intuitive literal response needs to be inhibited in order to correctly derive the implicatures in the reference game. Based on this intuition, we included the Stroop task (Golden & Freshwater, 1978), in which participants name the word color as quickly as possible while suppressing word semantics, as a measure of inhibition ability.

Cognitive Reflection Test (CRT) The Cognitive Reflection Test (Frederick, 2005) is a measure of how likely a person is to override their intuitive response and engage in further thinking – so in a sense, it is also a measure of inhibition ability, but while in the Stroop task, it is obvious what the correct answer should be, CRT contains "trick" questions where the answer isn't obvious, so it taps into a distinct concept.

Since CRT is know to be affected by familiarity (Stieger & Reips, 2016), we used a new CRT version with 6 critical questions, 3 verbal and 3 involving computation, and 4 non-trick "decoy" questions. The questions were selected from previously used versions of CRT (Primi, Morsanyi, Chiesi, Donati, and Hamilton (2016), Baron, Scott, Fincher, and Metz (2015), Sirota and Juanchich (2018), Thomson and Oppenheimer (2016), Toplak, West, and Stanovich (2014))[1] The items weere presented in randomized order in order to prevent participants from expecting to be tricked every time. Additionally, at the end of the experiment we asked the participants if they had seen any of the questions before.

The score is the proportion of correctly answered questions. Participants who had seen 3 or more questions before were excluded from analysis.

Experiments

Exp. 1. Pilot

We replicated Franke and Degen (2016)'s experiment and additionally collected a number of individual difference measures to investigate which of the individual differences predict the performance on the reasoning task.

Participants 95 native English speakers were recruited via Prolific.

Procedure Participants completed the experiment in three sessions. In the first session, they completed the reference game from Franke and Degen (2016), followed by a Need-for-Cognition Questionnaire. The second session took place a month later and contained Operation Span and the AQ. The third session took place a week after the second and contained Raven's Progressive Matrices, CRT, and the Stroop Task.

Results Following Franke and Degen (2016), only participants whose accuracy on the unambiguous trials was at least 95% entered analysis (for us it was fewer than 9% of the participants; Franke and Degen (2016) removed the bottom 15%). 8 of the 95 participants were excluded at this stage, with 87 participants remaining. Not all participants returned for the later experimental sessions; we also excluded one person who did not meet the 80% accuracy criterion for OSpan and 3 participants who indicated that they had seen some of the CRT questions but did not specify which ones. The number of participants whose data remained for each task after exclusions is reported in Table 1.

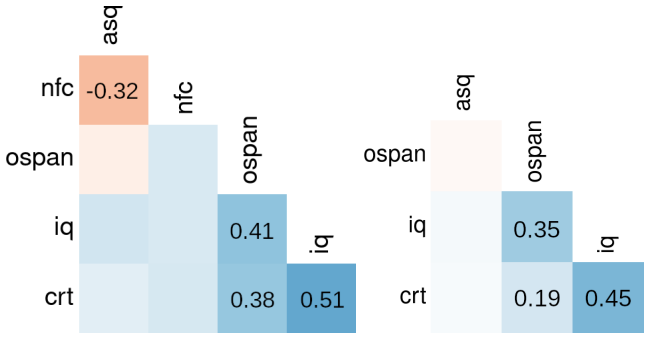
Means and ranges for the individual difference measures at a glance are reported in Table 1. The correlation matrix is shown in Figure 2. We see that IQ, CRT, and OSpan are all moderately positively correlated with each other, which is not surprising since all of these measures tap into aspects of reasoning ability. AQ is negatively correlated with NfC – that is, people who show more autistic traits judge themselves as less willing to engage in effortful cognitive activities. Stroop did not significantly correlate with any of the other measures.

Main task replication results Proportions of choice types per condition is displayed in Figure 3. They mirror the results from Franke and Degen (2016)'s original study very closely,

[1]The materials, as well as anonymized data and analysis scripts, are available at this link: `https://github.com/sashamayn/refgame_cogsci22`.

Table 1: Individual difference results summary for Exp 1, including the number of participants who completed each task after exclusions.

Measure	Mean	SD	Obs. range	Poss. range	*N*
NFC	3.37	0.9	1-5	1-5	87
OSpan	0.92	0.1	0.37-1	0-1	65
AQ	21.66	6.82	7-40	0-50	67
IQ	5.66	2.05	1-9	1-10	53
Stroop	0.49	0.25	0-1	0-1	51
CRT	0.41	0.27	0-1	0-1	50



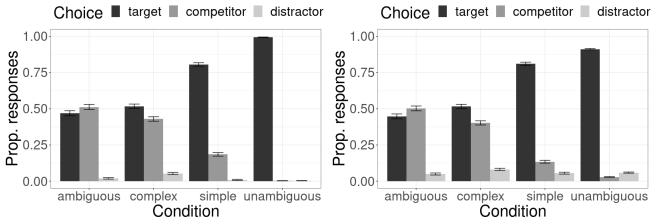
(a) Pilot (b) Confirmatory study

Figure 2: Correlation matrices for individual difference measures for the two experiments. Only the coefficients of significant correlations (corrected with Holm's method) are included. Stroop is omitted for Experiment 1 because it did not significantly correlate with anything.

with 81% of target and 19% of competitor responses for the simple condition for our study, and 77% and 23% for Franke and Degen (2016), and 52% of target and 43% of competitor responses for our pilot and 57% and 42% for the original study. Participants were at ceiling on the unambiguous trials and at chance on the ambiguous ones. The fact that participants' performance on the critical simple trials was better than on the complex ones suggests that the complex trials were indeed more difficult.

Following Franke and Degen (2016), we conducted logistic mixed-effects regression to verify the significance of the apparent differences between the simple and the complex condition, and between the complex condition and the chance baseline (ambiguous condition). The analysis setup and variable coding followed the original paper. The results once again match the ones from the original study very closely. As in the original study, participants performed significantly better on the simple trials than on the complex ones (β=1.52 (0.1), $p<0.0001$), while still performing above chance on the complex trials (β=0.3 (0.1), $p=0.004$). The only other significant predictor is the position of the target: participants chose the target more often if it was in the center (β=0.67 (0.11), $p<0.0001$) or on the right (β=0.35 (0.1), $p<0.001$).

In order to investigate the possible sources of individual



(a) Pilot (b) Confirmatory study

Figure 3: Choice proportions per condition for the two experiments. The results are quite similar for the two experiments and mirror those of the original study closely.

variation, we turned to a continuous analysis. To this end, we extended the minimal mixed effects model from the main task described above, and added individual difference measures to it. We removed the ambiguous trials from consideration since those only served as a baseline to show that participants performed above chance on the complex trials and have no relation to individual differences. The binary response correctness was regressed onto condition (now binary, simple vs. complex), target position, and main effects of the six individual differences. Like the original model, it included by-trial random slopes per participant. There are 47 participants for whom we have all individual measures. All individual difference measures were centered and rescaled to 0-1. The full model is reported in Table 2. Only the main effect of CRT is significant, suggesting that performance on this pragmatic reasoning task is modulated by the ability to override the intuitive response (presumably, the literal one or one dictated by salience) to complete the reasoning and make the correct prediction.

Since running the pilot across several sessions lead to a high dropout rate and a smaller number of participants than we had originally hoped for, we conduct an follow-up confirmatory study, very similar in structure but with a separate larger group of participants. This separate study will also serve to test whether the effect of CRT on pragmatic reasoning replicates. In order to make it feasible to collect the confirmatory study in a single session, we decided to drop two of the individual difference measures. Our analyses showed that NfC and the Stroop task were least promising, in that their effect sizes were very small in models where they were used as the only ID predictor (0.04 (0.59), $p=0.95$ for NfC and -0.23 (0.57), $p=0.69$ for Stroop). Therefore, these measures were dropped for the main experiment.

Exp 2. Confirmatory study

Participants 101 additional participants, all of them native English speakers, were recruited via Prolific.

Procedure The procedure was identical to that of the pilot, with the following differences. Participants completed the experiment in one session to avoid participant attrition. They

Table 2: Models for individual differences, based on the exploratory study (left), confirmatory study (middle), and the original study by Franke and Degen (2016) (right).

Fixed effect	Exp. 1 (47): $\beta(SE), p$	Exp. 2 (68): $\beta(SE), p$	F&D (51): $\beta(SE), p$
Intercept	0.05 (0.89), 0.95	-0.82 (0.51), 0.11	-0.15 (0.11), <0.18
condition (complex vs. simple)	**1.43 (0.16), <0.0001**	**1.87 (0.15), <0.0001**	**1.28 (0.12), <0.0001**
target position: middle vs. left	**0.79 (0.19), <0.0001**	0.13 (0.16), 0.41	**0.74 (0.13), <0.0001**
target position: right vs. left	0.29 (0.18), 0.1	0.13 (0.16), 0.41	**0.22 (0.08), <0.01**
NFC	-0.07 (0.56), 0.9	-	-
OSpan	0.2 (0.83), 0.81	-0.09 (0.51), 0.86	-
IQ	0.28 (0.6), 0.64	**1.75 (0.5), <0.0001**	-
Stroop	0.45 (0.55), 0.41	-	-
AQ	0.74 (0.61), 0.22	-0.58 (0.41), 0.16	-
CRT	**1.35 (0.56), 0.02**	**1.03 (0.39), 0.009**	-

Table 3: Individual difference results summary for Exp 2.

Measure	Mean	SD	Range
OSpan	0.93	0.11	0.38-1
IQ	5.69	2.12	0-10
AQ	20.5	7.54	7-42
CRT	0.43	0.26	0-1

completed the reference game and the four individual difference tasks in the following order: reference game, OSpan, CRT, AQ, and Raven's Progressive Matrices. Also, in the CRT, we now asked for each individual question whether participants had seen it before to avoid unnecessary exclusions.

Results We exclude participants who responded to fewer than 80% of the unambiguous filler items correctly, corresponding to 15% of the total participant number[2]. 4 further participants were excluded because they did not complete the experimental session. 12 further participants were excluded because they reported having seen 3 or more of the critical questions on the CRT. 5 further participants were excluded because their OSpan accuracy was under 80%. The remaining 68 participants entered the analysis. The means and ranges for the individual difference measures are reported in Table 3. They are similar to those we obtained in the pilot. We again observe significant correlations of CRT, OSpan, and IQ with each other, suggesting a shared reasoning component in these three measures (Figure 2, right panel).

The regression model without the individual difference replicated the same main effects as pilot. The model for the individual differences analysis can be found in Table 2. The model again revealed a significant main effect of CRT (β=1.03 (0.37), p=0.009). In addition, a significant effect of IQ (β=1.75 (0.5), p<0.0001) was revealed. There was no effect of AQ, suggesting that performance on this kind of pragmatic task is related to reasoning ability but not to socio-pragmatic personality traits. There was also no effect of OSpan.

We now turn to formal model predictions of individual differences. As can be seen in Figure 4, there is substantial variation in participant performance. Franke and Degen (2016) observed that a Bayesian hierarchical model which assigned one of three reasoning types (literal L_0, exhaustifier L_1, or Gricean L_2) to each participant obtained a better fit to the data than a homogeneous model that assumes one type of all the participants. We ran Latent Profile Analysis (LPA) using the *tidyLPA* package in R on our data for the main task.[3] LPA identifies clusters in the data while remaining agnostic to the origin of those clusters. The best fit is obtained with 4 classes (AIC = -82.9, BIC = -54.05); however, closer examination revealed that those 4 classes are not interpretable, and since we have a theoretical motivation to postulate 3 classes, we turn to the 3-class model (AIC = -71.75, BIC = -49.56), which had better fit than a homogeneous (AIC = -39.7, BIC = -30.81) or a 2-class (AIC = -41,84, BIC = -30.81) model.

We see that the 3 classes identified by LPA correspond approximately to the predictions of the models of idealized reasoning types (Figure 4). Class 1 is the smallest class which is not able to solve simple or complex implicatures, corresponding to L_0; Class 2 does above chance on the simple implicatures but not complex ones, corresponding to L_1, and Class 3 is able to solve both kinds of implicatures, corresponding to L_2.

Table 4 displays individual measures per class. Since only 4 people were assigned to Class 1, corresponding to the level-0 listener, the averages in that column should be interpreted with caution. However, when we compare Class 3 with Class 2, we observe a noticeable average increase in IQ and CRT, in line with the output of the mixed-effects model, as well as a slight increase in OSpan.

[2]In this confirmatory experiment, participants performed worse on the unambiguous fillers on average. Franke and Degen (2016)'s reason for having a strict exclusion criterion had to do with avoiding inflating the noise parameter when performing Bayesian model comparison. Our objective here is different – it is to investigate which individual differences predict performance in the reference game. We hence only exclude participants who aren't paying attention, and 80% on the fillers is a sufficiently high accuracy for our purposes.

[3]We use Model (1 equal variance, fixed covariance) since that is the model that makes the least assumptions and since some of the other models could not be estimated for our data.

Figure 4: Classes of participants identified by LPA for the reasoning task. The classes approximately correspond to the theoretical predictions of L_0, L_1, and L_2 respectively.

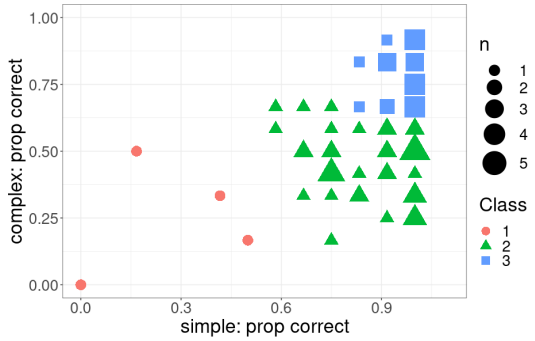


Table 4: Individual difference measure averages for the 3 classes predicted by LPA.

Measure	Class 1 (4)	Class 2 (41)	Class 3 (23)
Ospan	0.95 (0.04)	0.91 (0.14)	0.97 (0.05)
CRT	0.38 (0.25)	0.37 (0.24)	0.53 (0.28)
IQ	5 (2.94)	5.1 (1.97)	6.87 (1.79)
AQ	20.5 (10.66)	21.02 (7.36)	19.57 (7.59)

Discussion

While population-level models might provide a good fit to the data at the population level, they may also mask meaningful individual-level differences. We replicated Franke and Degen (2016)'s comprehension experiment, which revealed a large amount of individual variation in the rate at which people successfully derive implicatures, and replicated their finding according to which participants naturally fall into different groups corresponding to the theoretical predictions by the IBR models using an LPA analysis. We also collected individual difference measures of cognition and personality in order to learn more about the nature of the differences between these groups.

We found the Cognitive Reflection Test (CRT) to be a significant predictor of the ability to draw pragmatic inferences both in the pilot and in the confirmatory study. This measure reflects reasoning ability and the ability to override the first intuitive response. In our confirmatory study, we additionally find IQ to be a significant predictor of drawing inferences. We note that this effect did not turn out to be significant in the pilot, and hence needs to be interpreted with caution. We furthermore note that IQ is highly correlated with the cognitive reflection test (r=0.51 and r=0.46 for the two experiments) and it is hence unclear whether or not it contributes a real separate dimension. Future work could look into whether the effect of these two measures could be explained via one composite measure, or whether they represent two related but distinct abilities which influence pragmatic responding. For instance, Toplak, West, and Stanovich (2011) showed that CRT was able to account for some of the variance in performance on heuristic-and-bias tasks beyond measures of intelligence.

We did not observe an effect of socio-pragmatic abilities or working memory. However, we wonder whether the reason for the lack of a WM effect might lie in our participants' close-to-ceiling performance on the OSpan task (average score of 0.92 and 0.93 for the two experiments) – so perhaps there wasn't sufficient variability between participants to reveal an effect. This could be explored further by including more challenging working memory measures, e.g., OSpan with longer letter sequences.

It is interesting that only an effect of measures of reasoning ability was found, in contrast to some prior work on pragmatic comprehension of individual differences, such as Yang et al. (2018) who found an effect of both WMC and socio-pragmatic abilities (AQ) on sensitivity to context in scalar implicature derivation. It is possible that this is due to the nature of the reference game, which may involve explicit reasoning to a greater extent than other more naturalistic language tasks. It would be worth exploring in future research, therefore, to what extent the reasoning depth of an individual is task-dependent, i.e., would an L_2 type responder from this task also show evidence for sophisticated inferencing in other pragmatic tasks, including ones with a more prominent linguistic component like irony detection.

In general, probabilistic models of pragmatics like RSA and IBR are assumed to be computational-level and do not make explicit predictions of the processing mechanism of an n-level reasoner. It is possible, therefore, that several different mechanisms or strategies fall under the same probabilistic reasoning model. For instance, a closer look at the outlier assigned to Class 1, where accuracy is 0 for both implicature types revealed that this participant chose the distractor every time for both implicature types. Presumably, the participant reasoned that by sending e.g. the red hat, the speaker meant to communicate "the only creature *without* the red hat"; the participant did not pick randomly but applied a reasoning strategy, albeit not the correct one. In Mayn and Demberg (2022), we elicit participants' reasoning strategies for this task and find that, indeed, different participants use different strategies and that sometimes several reasoning strategies, some of which do not involve counterfactual reasoning but rather rely on e.g. surface-level similarity of the stimuli or salience, may result in the same surface-level performance.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. 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

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, *4*(3), 265–284.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of autism and developmental disorders*, *31*(1), 5–17.

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the raven's standard progressive matrices test. *Assessment*, *19*(3), 354–369.

Cohen, A. R., Stotland, E., & Wolfe, D. M. (1955). An experimental investigation of need for cognition. *The Journal of Abnormal and Social Psychology*, *51*(2), 291.

Fairchild, S., & Papafragou, A. (2021). The role of executive function and theory of mind in pragmatic computations. *Cognitive Science*, *45*(2), e12938.

Fairchild, S. C. (2018). *Speaker and listener effects on the processing of pragmatic meaning*. Unpublished doctoral dissertation, University of Delaware.

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.

Franke, M., et al. (2009). *Signal to act: Game theory in pragmatics*. Institute for Logic, Language and Computation Amsterdam.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*(4), 25–42.

Golden, C. J., & Freshwater, S. M. (1978). Stroop color and word test.

Heyman, T., & Schaeken, W. (2015). Some differences in some: examining variability in the interpretation of scalars using latent class analysis. *Psychologica Belgica*, *55*(1), 1.

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience*, *31*(1), 80–93.

Mayn, A., & Demberg, V. (2022). Right for the wrong reason? on the importance of eliciting participants' reasoning. *Under review*.

McVay, J. C., & Kane, M. J. (2012). Why does working memory capacity predict variation in reading comprehension? on the influence of mind wandering and executive attention. *Journal of experimental psychology: general*, *141*(2), 302.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (irt). *Journal of Behavioral Decision Making*, *29*(5), 453–469.

Scholman, M. C., Demberg, V., & Sanders, T. J. (2020). Individual differences in expecting coherence relations: Exploring the variability in sensitivity to contextual signals in discourse. *Discourse Processes*, *57*(10), 844–861.

Sirota, M., & Juanchich, M. (2018). Effect of response format on cognitive reflection: Validating a two-and four-option multiple choice question version of the cognitive reflection test. *Behavior research methods*, *50*(6), 2511–2522.

Stieger, S., & Reips, U.-D. (2016). A limitation of the cognitive reflection test: familiarity. *PeerJ*, *4*, e2395.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making*, *11*(1), 99.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition*, *39*(7), 1275–1289.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147–168.

Yang, X., Minai, U., & Fiorentino, R. (2018). Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in psychology*, *9*, 1720.