Title
What inferences do people actually make upon encountering informationally redundant utterances? An individual differences study

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Ryzhova, Margarita
Mayn, Alexandra
Demberg, Vera

Publication Date
2023

Peer reviewed

What inferences do people actually make upon encountering informationally redundant utterances? An individual differences study

Margarita Ryzhova, **Alexandra Mayn** and **Vera Demberg**
{mryzhova, amayn, vera}@lst.uni-saarland.de
Department of Language Science and Technology, Saarland University
66123 Saarbrücken, Germany

Abstract

Utterances mentioning a highly predictable event are known to elicit atypicality inferences (Kravtchenko and Demberg, 2015; 2022). In those studies, pragmatic inferences are measured based on typicality ratings. It is assumed that comprehenders notice the redundancy and "repair" the utterance informativity by inferring that the mentioned event is atypical for the referent, resulting in a lower typicality rating. However, the actual inferences that people make have never been elicited. We extend the original experimental design by asking participants to explain their ratings and administering several individual differences tests. This allows us to test (1) whether low ratings indeed correspond to the assumed inferences (they mostly do, but occasionally participants seem to make the inference but then reject it and give high ratings), and (2) whether the tendency to make atypicality inferences is modulated by cognitive factors. We find that people with higher reasoning abilities are more likely to draw inferences.

Keywords: pragmatics; individual differences; atypicality inferences; script knowledge; informational redundancy; reasoning

Introduction

According to Grice (1975), overinformativeness or informational redundancy relates to the violation of the Quantity Maxim and should be avoided in rational communication. However, overinformative utterances have been found to be surprisingly common (Rubio-Fernández, 2016; Degen, Hawkins, Graf, Kreiss, & Goodman, 2020; L. R. Horn, 1991; L. Horn, 1993, 2014). In turn, listeners have been found to respond to overinformative messages differently: e.g., by either tolerating them (Davies & Katsos, 2010), exploiting the redundant information for more effective visual search (Rubio-Fernandez, 2019), or generating inferences (Rohde, Futrell, & Lucas, 2021; Kravtchenko & Demberg, 2022).

We here specifically focus on the atypicality inferences investigated first in Kravtchenko and Demberg (2015), and later replicated in (Ryzhova & Demberg, 2020; Ryzhova, Loy, & Demberg, 2022; Kravtchenko & Demberg, 2022). They studied utterances that are informationally redundant on the basis of script knowledge:

(1) *Mary went to a restaurant.* ***She ate there!***

In the above example, the first sentence refers to an activity that is part of the common sense knowledge of most people (i.e., about what happens when you go to a restaurant such as being seated, ordering, eating, paying). In the literature, scripts are defined as knowledge structures built on past experiences about common routine or conventional activities (Schank & Abelson, 1975). It has been shown that events constituting a script normally do not need to be mentioned explicitly but are automatically inferred by the comprehender on the basis of the script topic itself (Zwaan, Magliano, & Graesser, 1995; Bower, Black, & Turner, 1979). In this sense, the utterance "*She ate there!*" from example (1) is considered informationally redundant (IR) because eating is an integral part of *going to a restaurant* script and is automatically inferred by the reader.

Kravtchenko and Demberg (2015) showed that mentioning such events overtly triggers pragmatic inferences. When comprehenders encounter such an IR utterance, they rate the probability of Mary usually eating in a restaurant lower than if it is not mentioned. Such belief changes were interpreted as a repair mechanism that accommodates the common ground to make the IR utterance informative with respect to the context (e.g., that Mary does not usually eat at restaurants). Consequently, the behavior that is typically entailed by the context (eating when going to a restaurant) becomes atypical for the utterance's referent. Atypicality inferences are highly context-sensitive and were shown to disappear in atypical contexts (e.g., when Mary is described as a person who does not like eating out) or when the target utterance refers to an event that is not highly associated with an everyday activity (e.g., "*Mary went to a restaurant.* ***She got to see their kitchen!***").

Atypicality inference may consist of several steps: first, the informational redundancy needs to be noticed, and then an accommodation process needs to occur. It results in the inference being drawn that the mentioned event is **not** predictable for the person in question (i.e., Mary doesn't usually eat in restaurants). This would then be reflected in a lower event typicality rating. One can also assume that in order to obtain a consistent picture, the comprehender would maybe even come up with a reason that could lead to the event being worth mentioning (e.g., it is not redundant to say that Mary ate in the restaurant because Mary typically goes to restaurants for drinks but doesn't order food). However, it has never been tested directly whether participants indeed make such inferences. In general, we expect lower ratings to correspond to atypicality inferences and higher ratings to the lack thereof; however, ratings may also be noisy and may mask certain underlying processes. For example, if subjects first compute the

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

implicature but then reject it, it won't be reflected in the typicality ratings, as those only show participants' final decision.

In the following study, we address this question by asking participants to provide an explanation for their typicality estimates of the target event (i.e., how often they think Mary usually eats when going to a restaurant and why). Annotating these explanations gives us a qualitative and quantitative picture of how people accommodate redundancy and allows us to check whether the atypicality interpretations are indeed identifiable based on the typicality ratings.[1].

We were also interested in whether participants' biases in their interpretation of informational redundancy are consistent and, if so, whether they are modulated by cognitive or personality traits. Previous research on pragmatic processing has already shown for scalar inferences that are quite consistent within a participant (Heyman & Schaeken, 2015), and subsequent work has provided further evidence for the existence of individual variability in processing different pragmatic phenomena (e.g.,Antoniou, Cummins, and Katsos (2016); Fairchild and Papafragou (2021); Yang, Minai, and Fiorentino (2018)).

To date, the research on individual differences in pragmatic processing has been restricted to generalized implicatures, while the processing of particularized implicatures, such as atypicality inferences, has been less in focus.

Given that previous studies on atypicality inferences used a few-shot approach, it is unclear to what extent the drawing of the inference is consistent within a specific participant and whether someone's tendency to draw or not draw such inferences can be explained by their individual cognitive traits (Ryzhova et al., 2022). In the next section, we discuss what measures might modulate the processing of atypicality inferences based on previous work.

Individual differences in pragmatic processing and atypicality inferences

We collected the following individual difference measures to investigate whether they predict participants' responding tendencies in the atypicality inference task.

Verbal working memory capacity It has been argued that implicature derivation is effortful and therefore requires sufficient cognitive resources (Antoniou et al., 2016; De Neys & Schaeken, 2007; Fairchild & Papafragou, 2021). For example, Yang et al. (2018) found that individuals with higher working memory capacity showed higher context sensitivity when deriving scalar implicatures. We hypothesized that atypicality inferences may also be costly and draw on executive function resources – meaning that individuals with higher working memory capacity would be more likely to derive them.

Verbal working memory capacity was measured using the Reading Span task (Caplan & Waters, 1999; Scholman, Demberg, & Sanders, 2020).

[1]Data and analysis code are available at `https://github.com/mnryzh/atypicality_inddiff`

Cognitive reflection It could be that derivation of some inferences requires overriding the literal interpretation to arrive at the pragmatic one. The Cognitive Reflection test (Frederick, 2005) taps into reflexivity and the tendency to override the intuitive but wrong response (Pennycook, Cheyne, Koehler, & Fugelsang, 2016; Welsh, 2022). The rate of pragmatic responding in a reference game was shown to be modulated by participants' performance on the Cognitive Reflection Test (Mayn & Demberg, 2022), while Heyman and Schaeken (2015) found that participants with higher CRT scores were more consistent in their responses to underinformative sentences. If atypicality inferences behave similarly, individuals with higher ability to override the intuitive response would be more likely to derive them.

We used a 10-question version of CRT, with 6 critical questions, 3 verbal and 3 computational, and 4 decoy questions 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)), presented in random order. Since CRT is known to be affected by familiarity (Stieger & Reips, 2016), we asked the participants after each question whether they had seen it before, and computed the score as the proportion of correctly answered previously unseen critical questions. Participants who reported having seen 3 or more of the 6 critical questions were excluded from the analysis.

Exposure to print and language experience It has been shown that individuals with higher print exposure are more sensitive to certain context cues (Arnold, Strangmann, Hwang, Zerkle, & Nappa, 2018; Scholman et al., 2020). We hypothesized that individuals with higher print exposure might notice and react to informational redundancy more easily, which would make them more likely to derive atypicality inferences.

We used the Author Recognition Test (ART) (Stanovich & West, 1989; Acheson, Wells, & MacDonald, 2008; Martin-Chang & Gould, 2008) to measure print exposure, where participants are asked to recognize authors in a list of names, where half of the names are real authors and the other half are foils.

Non-verbal intelligence In order to derive an atypicality inference, participants need to reason about the possible contexts in which the apparently redundant utterance may not be redundant anymore (e.g., stating that Mary ate at a restaurant is not redundant if she usually only orders drinks). We hypothesized that abstract reasoning ability may modulate the process of coming up with a context that would accommodate the atypicality inference. Also, a positive effect of reasoning ability on pragmatic responding was observed in a pragmatic reference game (Mayn & Demberg, 2022).

Raven's Progressive Matrices Test (IQ) (Raven, Raven, & Court, 1962) was included as a test of nonverbal intelligence. 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.

Socio-pragmatic abilities, as measured by the Autism Spectrum Quotient, have been found to correlate with pragmatic responding (Yang et al., 2018): people who are higher in autism may be less likely to put themselves in the interlocutor's position or reason about why the interlocutor said what they did, therefore responding more literally. Similarly, Nieuwland, Ditman, and Kuperberg (2010) found that underinformative sentences elicited an N400-effect only in pragmatically skilled participants, as indicated by low scores on the AQ Communication Subscale. Adults diagnosed with ASD have been shown to perform significantly worse on tests of Theory of Mind (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Happé, 1994). We hypothesized that people with higher scores on the AQ (indicating more autistic tendencies) would be less likely to draw an atypicality inference as they might have more trouble recognizing the redundancy and inferring its nonliteral meaning. The Autism Spectrum Quotient (AQ) (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) consists of 50 statements which are meant to tap into autistic traits.

Experiment

Participants

336 subjects were recruited via the online crowdsourcing platform Prolific. All participants were native speakers of English with an approval rating of at least 95%.

228 of the 336 subjects returned to participate in the second session[2]. 11 subjects had to be excluded due to data loss on the server. 24 subjects were excluded from analyses due to exclusion criteria of one of the individual measures (19 subjects were familiar with 3 or more critical questions on the CRT, and 5 subjects always or almost always responded that they did not know the author on the ART[3]). The remaining 193 participants entered the analyses.

Materials

We used the materials from Kravtchenko and Demberg (2015), which consist of twenty-four brief stories describing different everyday situations, such as grocery shopping or going to a restaurant (see an example of an item in Table 1). Every item includes a context of a few sentences introducing the topic and story characters. The continuation of a story depends on the condition. Critical items contain an utterance by one of the characters stating that an activity that is highly predictable in the given context took place (with-IR condition – with informational redundancy). The IR-utterances are marked with an exclamation mark. We note that Kravtchenko and Demberg (2022) showed that atypicality inferences do not hinge on the exclamation mark – they also occur, albeit with a smaller effect size, if the IR utterance is marked only with a full stop. For increased power in our individual difference study here, we decided for stimuli including the exclamation mark.

Control items were identical to the critical items except they did not contain the redundant utterance (without-IR condition). Filler items contained an utterance that was not informationally redundant given the context.

There were two questions following each item. In the first question, participants were asked to provide the typicality of the target activity on a scale of 0 ("Never") to 100 ("Always") using a slider. For the example in Table 1, the first question would thus be "*How often do you think Mary eats at restaurants when she goes there?*". Participants needed to click on the scale for the slider to appear, to make sure that they didn't just click through, leaving it in the initial position. On filler trials, half of the questions were about an event that was not mentioned in the utterance, such as, "How often do you think Mary gets to see the kitchen when she goes to restaurants?" in order to mask the experimental manipulation.

When participants gave a rating and clicked on "Next question", the slider froze and a textbox appeared, along with the question "*Why did you put the slider in this particular position?*". This was done as a probe into participants' reasoning and to investigate whether a low rating indeed corresponded to an atypicality rating triggered by the informationally redundant utterance.

Experimental procedure

In order to keep the original design of showing each participant only very few items and thus avoid possible learning effects (Kravtchenko & Demberg, 2015, 2022), but at the same time have more data per subject for our analyses, we conducted the main experiment in two experimental sessions with at least two weeks in between.

The 24 items from Kravtchenko and Demberg (2015) were used to construct 8 balanced experimental lists with 3 items appearing in the target with-IR condition, 3 in the control without-IR condition, and 4 in the filler condition in each list. In the second session, the subjects saw lists of the same structure but consisting of items they had not seen in the first session. Across the two sessions, we thus obtained 6 observations per subject in the with-IR condition and 6 in the without-IR condition.

In the second session, all participants additionally completed a battery of five cognitive and personality tests in the following order: Reading Span test (RSpan), Cognitive Reflection Test (CRT), Author Recognition Test (ART), Raven's Progressive Matrices Test (IQ), and Autism Spectrum Quotient (AQ).

Annotation procedure

Participants' explanations of why they assigned a particular typicality rating were annotated by two raters using the labels shown in Table 2. The annotators only saw the textual

[2]Some participants were not informed about the two rounds of data collection, which resulted in a lower retention rate.

[3]We only excluded subjects who consistently responded that they did not know the author even for very famous authors.

Table 1: An example of "going to a restaurant" story in with-IR, without-IR, and filler conditions

context	Mary is a journalist who often goes to restaurants after her interviews. Yesterday she went to a popular Chinese place where she ran into her friend David. Later that day David ran into Sally, a mutual friend of him and Mary.		
condition	**with-IR**	**without-IR**	**filler**
target activity	David said to Sally: "I ran into Mary leaving that Chinese place. She ate there!"	–	David said to Sally: "I ran into Mary leaving that Chinese place. She recently got a promotion!"
Question 1	How often do you think Mary usually eats, when going to a restaurant?		How often do you think Mary usually gets to see the kitchen, when going to a restaurant?
Question 2	Why did you place the slider in this particular position?		

responses but not the typicality ratings provided in the first question, so as to avoid possible bias.

Subjects' responses were classified as *normal* if they stated that the subject was likely to have performed the predictable activity as it is normal in the given context. This corresponds to a participant not making a pragmatic inference. The tag *atypicality* corresponds to the participant making an atypicality inference. For example, an explanation like "Since it was mentioned explicitly, maybe sometimes Mary does not eat at restaurants and just orders a drink" would be labeled *atypicality*. We also found some instances where the participant indicated that they had made the atypicality inference, but did not accept this inference, see Table 2 for an example. We annotated these instances as *notice_reject*. If the participant reported being unsure, their response was labeled *not_sure*. Finally, responses that did not fall into any of the above categories were labeled *other*. This tag was used if multiple of the above tags could be applicable to the provided explanation or if it was completely unclear what the participant meant. Example responses for each tag for the restaurant story are reported in Table 2.

Table 2: Examples of the annotations from the restaurant item

annotation tag	**inference drawn**	**example**
normal	no	Usually when you go to a restaurant, it is to eat.
atypicality	yes	Since David mentioned it, it sounds like she doesn't always eat at restaurants. Maybe she sometimes interviews people in restaurants.
notice_reject	unclear	After interviews Mary will be tired so she probably eats. She can't just go to a restaurant for a drink after a long day.
not_sure	unclear	I'm not sure.
other	unclear	He didn't tell Sally which restaurant, he said that restaurant, as though they go there often.

The inter-annotator agreement was substantial (Cohen's κ = 0.74 ($p<.0001$), 95% CI (0.7, 0.77)). All disagreements were resolved jointly.

Results

Individual measures Descriptive statistics of the individual difference measures are reported in Table 3, and the correlations are reported in Figure 1. Because some of the measures were correlated, with the highest correlation being between CRT and IQ at r=0.36, we performed PCA to see whether any of the individual measures load onto the same component. Indeed, the five individual differences were best explained by four components, with IQ and CRT loading onto the same component. Therefore, in our models, we use a single composite score for these two measures, which we call the Reasoning score.

Table 3: Descriptive statistics for the individual differences.

Task	Poss. range	Obs. Range	Mean (SD)
ART	-65 – 65	-9 – 58	19.3 (14.28)
AQ	0 – 50	2 – 49	20.52 (8.37)
CRT	0 – 1	0 – 1	.31 (.26)
IQ	0 – 10	1 – 10	5.38 (2.12)
RSpan	0 – 1	.01 – 1	.76 (.2)

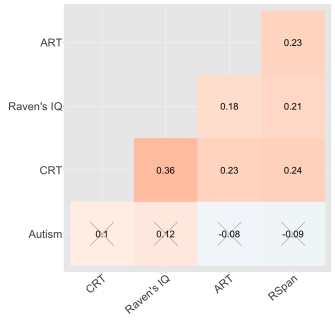

Figure 1: Correlations between the individual difference measures (non-significant correlations are crossed-out)

Analysis of ratings We built a beta mixed effects regression model of participants' ratings in order to replicate the

main pragmatic effect (lower rating in the with-IR condition). The ratings were transformed to fit a beta distribution[4] and regressed onto the story condition (with-IR vs. without-IR; ±0.5 sum-coded) and onto main effects of the four rescaled and centered cognitive measures and the interaction of the measures with story condition. Since the data was collected in two experimental sessions, we also added the session as a predictor. The random structure included by-subject and by-item random intercepts and by-item random slopes for the story condition[5]. The minimal model with only the significant predictors is presented in Table 4. We replicate the main pragmatic effect ($b = -0.45, z = -7.89, p < .001$): participants gave lower typicality ratings in the with-IR condition where the predictable event was explicitly mentioned ($mean = 72.63, sd = 29.82$) compared to the without-IR condition ($mean = 85.71, sd = 20.61$).

We also find a main effect of ART ($b = 0.08, z = 2.13, p = .03$), suggesting that participants with more reading experience generally give slightly higher ratings in both conditions. Also, we find an interaction of the AQ with condition ($b = -0.10, z = -2.07, p = .04$), where participants with higher scores on the AQ give lower ratings in the target with-IR condition. The direction of the effect is the opposite of what we had predicted: we expected people with higher AQ scores to be more literal and therefore less likely to make an atypicality inference. Finally, there is a trend of interaction between condition and reasoning ($b = -0.05, z = -0.51, p = .07$), suggesting that people with higher reasoning ability might be more likely to make an atypicality inference. None of the other effects reached significance, so they were not included in the final model.

We hypothesize that ratings, while being a good proxy for atypicality inferences (main effect of condition), may be too noisy for investigating the relationship between atypicality inferences and individual differences. Therefore, we next turn to annotations of participants' explanations as another possible proxy.

Analysis of annotations Figure 2 shows the mean rating associated with each annotation category (top panel) and the frequency of each tag (bottom panel). We see that *atypicality* is the most frequent tag (*N=528*), and it indeed corresponds to a much lower average typicality rating (*mean=51.84, sd=28.06*) than ratings given by people whose answer indicated that they made no atypicalty inference (*normal*) (*N=457, mean=93.82, sd=11.38*). Interestingly, the *notice_reject* cases were found to correspond to similar ratings as in the *normal* cases where the pragmatic inference was not made (*N=71, mean=95.46, sd =8.97*). While these results may not be surprising per se, they provide evidence that the ratings correspond well to comprehenders' inferences, and can validly be used as a proxy for whether atypicality inferences are made.

Table 4: Effect sizes (b), standard errors (SE), z-values, and p-values for the minimal mixed effects beta regression model of participants' ratings of the target activity typicality (the ratings were transformed to fit a beta distribution). The dispersion parameter is 1.87.

	b	*SE*	*z*	*p*
Intercept	1.35	0.07	18.99	<.001
Condition (with-IR)	-0.45	0.06	-7.89	**<.001**
AQ	-0.06	0.04	-1.64	.10
ART	0.08	0.04	2.13	**.03**
Reasoning	-0.01	0.02	-0.51	.61
Condition : Reasoning	-0.05	0.03	-1.82	.07
Condition : AQ	-0.10	0.05	-2.07	**.04**
Random effects	Variance			
Subject	0.16			
Item	0.08			
Condition | Item	0.03			

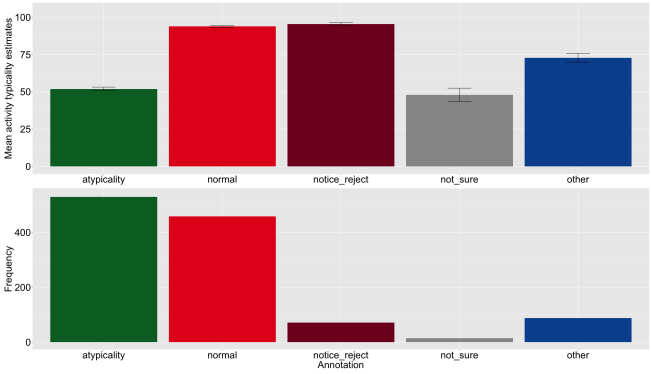

Figure 2: With-IR condition. Mean typicality ratings (±*SE*; upper panel) and counts (lower panel) per annotation tag.

We next take a look at the individual variability in participants' responding strategy by grouping subjects into three classes: *literal* or *pragmatic* if 4 or more of their 6 explanations in the with-IR condition were *normal*/*notice_reject* or *atypicality* respectively, and *inconsistent* otherwise. We found that the majority of participants showed consistent behaviour, falling either into the pragmatic class (*N=80*) or the literal class (*N=51*). For the pragmatic class, there is a sizeable difference in ratings between conditions (84.57 (21.28) in without-IR vs. 50.51 (30.49) in with-IR), whereas for the literal class there is no difference in typicality ratings (87.63 (18.76) in without-IR vs. 88.75 (20.20) with-IR condition).

Individual differences analysis We now examine how the likelihood of a participant giving a pragmatic response relates to their cognitive properties. We conducted a mixed effects logistic regression analysis where the dependent measure is

[4]The choice of beta distribution is justified by the nature of the ratings: the ratings are bounded by the experimental design (slider end points), and they exhibit a strong negative skew.

[5]The maximal random effects structure (that also included by-subject random slopes) was simplified to reach convergence

whether a participant's explanation in the with-IR condition made an atypicality inference, as evidenced by the *atypicality* tag (1 - atypicality inference was made, 0 - not). For this analysis, we coded *notice_reject* category as 0 (atypicality inference not made) and excluded the *not_sure* and *other* trials since for those trials it is unclear whether the subjects made an atypicality inference (8% of trials). The dependent measure was regressed onto the individual difference measures, as well as the session in which the participant saw a given item. The model also included per-participant and per-item random intercepts. We report the minimal model with only the significant predictors. The results of the minimal model are shown in Table 5. The model reveals a significant effect of reasoning (*b=0.23, SE=0.07, p=.001*), where participants with higher scores on IQ and CRT tend to give more pragmatic responses. Note that in the beta regression model of the ratings included this effect as a trend; here, however, it is highly significant. There is also a significant effect of AQ (*b=0.25, SE=0.12, p=.03*), again in the opposite direction from what we would expect: people who have higher AQ scores appear to give more pragmatic responses.

Table 5: Effect sizes (*b*), standard error (SE), z-values, and p-values for the minimal logistic regression model of the annotations of participants' explanations (atypicality vs. not atypicality) in the with-IR condition.

	b	*SE*	*z*	*p*
Intercept	-0.02	0.18	-0.14	.89
AQ	0.25	0.12	2.17	**.03**
Reasoning	0.23	0.07	3.22	**.001**
Random effects	Variance			
Subject	1.46			
Item	0.45			

Discussion

In this study, we were able to largely confirm the previous assumption that the reduction in typicality ratings for the redundant activity corresponds to an atypicality inference – subjects explained their belief changes by assuming that the target event, such as eating in the context of going to the restaurant, is not typical for the actor. We also find substantial variability in subjects' strategies of processing redundancy, with most participants either consistently drawing or consistently not drawing atypicality inferences.

Next, we related observed variability in subjects to their cognitive traits. We found an effect of reasoning (composite score of IQ and CRT) in the model with atypicality annotations and as a trend in the model with ratings, suggesting that subjects with higher reasoning ability draw more atypicality inferences. We hypothesize that this effect is driven by greater capacity to accommodate observed redundancy and, in particular, to come up with explanations of the apparently redundant utterance that would render it informative. This hypothesis is also in line with previous studies showing that people with higher cognitive reflection and fluid intelligence are more likely to engage in deeper processing as opposed to following cognitively inexpensive heuristics (Toplak, West, & Stanovich, 2011; Otero, Salgado, & Moscoso, 2022; Shtulman & Young, 2022). However, we note that this is not a direct connection to pragmatic processing. Neither IQ nor CRT, to our knowledge, has previously been used as measures of subjects' variability in pragmatic tasks (the only exception being Mayn and Demberg (2022), who found similar effects for pragmatic inferences in reference games). That is why, to be able to make more global conclusions about how reasoning might be related to pragmatic processing, other types of inferences should be tested in future work.

Our models also revealed an effect of AQ: people with higher AQ scores tended to draw more atypicality inferences. This finding is surprising and does not align with findings in the scalar implicatures, where more pragmatic answers were associated with less "autistic" profiles (e.g., see Yang et al. (2018)). Following previous literature, we used the total AQ score as the model predictor. However, there is a debate in the literature about what exactly the total AQ score measures, and there are recommendations to use subscales instead (English, Gignac, Visser, Whitehouse, & Maybery, 2020). When we replaced the total AQ score with the Communication/Mindreading subscale, which appears most relevant, the effect ceased to be significant. Therefore, replication and further investigation into what the AQ measures and how it is related to atypicality inferences is warranted.

We also found that, compared to ratings, annotations yield themselves better to statistical analysis of individual differences, as they represent a less noisy picture of what inferences were actually made. In particular, we observe a category of answers, $notice\_reject$, where participants went through the process of deriving the atypicality inference but then rejected it and gave high typicality ratings. Therefore, based on the ratings alone, those responses are indistinguishable from *normal*, where the atypicality inference was not even considered. In a post-hoc analysis, we therefore re-ran the regression model for ratings but excluded $notice\_reject$ instances from the analysis. We found that the interaction between reasoning abilities and condition in that case comes out more clearly ($b = -0.06$, $SE = 0.03$, $z = -2.04$, $p = .04$), supporting the important role of reasoning in drawing atypicality inferences. If we only consider the ratings, however, we will not be able to distinguish $notice\_reject$ responses from *normal* responses. Future work should hence collect both ratings and explanations or develop novel measures that can distinguish these cases more clearly.

Thus, annotations of people's explanations give us useful insights into how they accommodate redundancy. However, since these explanations are given after the fact, we cannot be sure that they perfectly match the thinking in the moment. In future work, it would be interesting to see how the processing of atypicality inferences is reflected in online measures.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments.

The first author was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Funder Id: https://doi.org/10.13039/501100001659, Project-ID 232722074 – SFB1102: "Information Density and Linguistic Encoding".

The other authors were funded by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878).

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior research methods*, *40*(1), 278–289.

Antoniou, K., Cummins, C., & Katsos, N. (2016). Why only some adults reject under-informative utterances. *Journal of Pragmatics*, *99*, 78–95.

Arnold, J. E., Strangmann, I. M., Hwang, H., Zerkle, S., & Nappa, R. (2018). Linguistic experience affects pronoun interpretation. *Journal of Memory and Language*, *102*, 41–54.

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., Hill, J., Raste, Y., & Plumb, I. (2001). The "reading the mind in the eyes" test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241–251.

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.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive psychology*, *11*(2), 177–220.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and brain Sciences*, *22*(1), 77–94.

Davies, C., & Katsos, N. (2010). Over-informative children: Production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua*, *120*(8), 1956–1972.

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.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: dual task impact on scalar implicature. *Experimental psychology*, *54*(2), 128.

English, M. C., Gignac, G. E., Visser, T. A., Whitehouse, A. J., & Maybery, M. T. (2020). A comprehensive psychometric analysis of autism-spectrum quotient factor models using two large samples: Model recommendations and the influence of divergent traits on total-scale scores. *Autism Research*, *13*(1), 45–60.

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

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

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, *24*(2), 129–154.

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.

Horn, L. (1993). Economy and redundancy in a dualistic model of natural language. *Sky*, *1993*, 33–72.

Horn, L. (2014). Information structure and the landscape of (non-) at-issue meaning. In *The oxford handbook of information structure.*

Horn, L. R. (1991). Given as new: When redundant affirmation isn't. *Journal of Pragmatics*, *15*(4), 313–336.

Kravtchenko, E., & Demberg, V. (2015). Semantically underinformative utterances trigger pragmatic inferences. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 37).

Kravtchenko, E., & Demberg, V. (2022). Informationally redundant utterances elicit pragmatic inferences. *Cognition*, *225*, 105159.

Martin-Chang, S. L., & Gould, O. N. (2008). Revisiting print exposure: Exploring differential links to vocabulary, comprehension and reading rate. *Journal of Research in Reading*, *31*(3), 273–284.

Mayn, A., & Demberg, V. (2022). Individual differences in a pragmatic reference game. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of memory and language*, *63*(3), 324–346.

Otero, I., Salgado, J. F., & Moscoso, S. (2022). Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. *Intelligence*, *90*, 101614.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang,

J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior research methods*, *48*, 341–348.

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.

Raven, J. C., Raven, J. C., & Court, J. H. (1962). *Advanced progressive matrices*. HK Lewis London.

Rohde, H., Futrell, R., & Lucas, C. G. (2021). What's new? a comprehension bias in favor of informativity. *Cognition*, *209*, 104491.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in psychology*, *7*, 153.

Rubio-Fernandez, P. (2019). Overinformative speakers are cooperative: Revisiting the gricean maxim of quantity. *Cognitive science*, *43*(11), e12797.

Ryzhova, M., & Demberg, V. (2020). Processing particularized pragmatic inferences under load. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 42).

Ryzhova, M., Loy, J., & Demberg, V. (2022). Pragmatic comprehension of implicatures—consistency within individuals across types and time. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Schank, R. C., & Abelson, R. P. (1975). Scripts, plans, and knowledge. In *Ijcai* (Vol. 75, pp. 151–157).

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.

Shtulman, A., & Young, A. G. (2022). The development of cognitive reflection. *Child Development Perspectives*.

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.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading research quarterly*, 402–433.

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.

Welsh, M. B. (2022). What is the CRT? intelligence, personality, decision style or attention? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

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

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of experimental psychology: Learning, memory, and cognition*, *21*(2), 386.