Individual differences in the delay effect across scales

Sonia Ramotowska (s.ramotowska@uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam

Abstract

For more than two decades, researchers have been trying to explain the source of the processing cost of scalar implicature (SI). Although the computation of some SIs is associated with longer processing time (known as the delay effect), other SIs are processed cost-free. In this study, we investigated how individual differences in the rate of SI derivation modulate the delay effect across different scales. We reanalyzed four datasets from two SI verification task studies, which examined various scales. In these experiments, participants judged SI-triggering sentences as either *true* (literal reading) or *false* (SI reading). We fit a computational model to quantify the by-subject probability of computing SIs. Across datasets, we found that subjects who prefer the literal reading of the SI-triggering sentence were faster to respond *true* than *false*. However, the reading preferences modulate the verification speed differently for different scales. This suggests that the source of the delay effect might vary between scales.

Keywords: scalar implicatures; scalar diversity; delay effect; individual differences

Implicatures and delay effect

The utterance "My phone battery is low" triggers certain inferences called implicatures (Grice, 1975). For example, the speaker may indirectly ask the listener to lend him a charger. He may also imply that "the battery is low but not empty", and thus the phone is still turned on. This second inference, scalar implicature (SI), arises when the recipient of the utterance compares what the speaker said ("My phone battery is low") to what he could have said ("My phone battery is empty"). This comparison is possible because the scalar words *empty* and *low* belong to one scale, and the former one (stronger scale mate) is more informative than the latter (weaker scale mate) (Horn, 1989). To draw SI, the listener must conclude that the speaker chose to utter the less informative sentence because he does not believe in the more informative one (cf. Grice, 1975). Thus, the speaker meant that "My phone battery is low but not empty".

We will refer to sentences including weaker scale mates as the *SI-triggering sentences*. The abundance of studies (see Khorsheed & Gotzner, 2023 for a review) have investigated in which context what kind of SIs arise, as well as how often and fast individuals draw implicatures. These studies often used a verification paradigm in which the SI-triggering sentence is presented in an ambiguous context. It can be judged as *true* according to its literal meaning or *false* according to its SI interpretation. The task allows measuring how frequently and fast implicatures are computed.

The classical debate between the Default view (Levinson, 2000) and the Relevance theory (Sperber & Wilson, 1986, 1987) concerns whether the computation of SIs involves a processing cost. According to the former view, SIs are generated without the processing cost. The latter view, in turn, predicts the processing cost because the sentence's literal meaning is always computed first, and only if the context supports the stronger interpretation, the SI is generated. Bott and Noveck (2004) tested the predictions of these two accounts and found support for the Relevance theory. They tested ambiguous sentences involving a scalar *some* (e.g., "Some elephants are mammals") that could be judged as *true* (literal response) or *false* (pragmatic response). They found the delay effect, meaning that participants who follow the literal interpretation are faster than participants who follow the SI interpretation.

Since this first study, multiple experiments have tested under which conditions the delay effect arises and when it is reduced (see e.g., Khorsheed, Price, & van Tiel, 2022 for review). However, several findings questioned the generality of the delay effect. Firstly, some studies showed that this effect does not generalize to all scales (Van Tiel, Pankratz, & Sun, 2019; Van Tiel & Pankratz, 2021). Secondly, it depends on individual preferences for the literal or SI reading (Kursat & Degen, 2020). Thus, to better understand whether the SI computation generates the processing cost, scalar diversity (Van Tiel, Van Miltenburg, Zevakhina, & Geurts, 2016) and individual differences must be considered.

Scalar diversity

Scalar diversity refers to the phenomenon that SIs are derived at different rates depending on the scale. Typically, studies investigating scalar diversity focus on linguistic properties of scales that change the rate of SI endorsements, such as the semantic distance between scale mates and the boundedness of the scales (Van Tiel et al., 2016), inferences drawn from the stronger scale mate (Gotzner, Solt, & Benz, 2018), or the accessibility of the stronger scale mate (Ronai & Xiang, 2022). In a verification task, Van Tiel et al. (2019) found that the rate of literal interpretations varies across scales from 38% for *some* to 71% for *low*. By correlating the number of *true* responses between scales, they showed that participants are not consistent in response choice between scales. Hierarchical cluster analysis revealed that the scalars *low, scarce* and

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

try constitute one cluster, and the scalars *might, some, or*, and *most* constitute the second cluster.

Not only do different scales give rise to different rates of SIs, but also the processing cost depends on the scale. Van Tiel and Pankratz (2021) and Van Tiel et al. (2019) showed that the polarity of the scale predicts the delay effect. Van Tiel et al. (2019) tested 7 different scalars and found the delay effect for positive scalars *or, might, most*, while a negative scalar *scarce* leads to the reversed effect, meaning that literal responses are slower than pragmatic responses (in the case of *some, try, low* there was no delay effect). Van Tiel and Pankratz (2021) expanded the number of tested scales and introduced a more nuanced measure of the polarity of the scale. They found that all six positive scales (with scalars *content, fair, passable, ajar, chubby, warm*) led to the processing cost of SI and only one negative scale (*youthful*). In general, both studies were interpreted as supporting the polarity explanation of the delay effect, according to which only positive scales are linked with the processing cost.

Individual differences

Already Bott and Noveck (2004) observed that the choice of SI interpretation is subject to individual differences. While literal responders judge SI-triggering sentences as *true*, pragmatic responders choose to respond *false*. Many studies have tried to explain this variability by various cognitive and personality factors (e.g., Antoniou, Cummins, & Katsos, 2016; Dieussaert, Verkerk, Gillard, & Schaeken, 2011; Fairchild & Papafragou, 2021; Feeney & Bonnefon, 2013). Moreover, individual differences in reading preferences affect the SI derivation speed. Heyman and Schaeken (2015) found that the consistently literal responders were faster than the consistently pragmatic responders or the inconsistent responders. Kursat and Degen (2020) showed that responders (literal or pragmatic) respond faster when responding consistently to their preferences. These studies, although limited to scalar *some*, suggest that the speed with which participants verify SI-triggering sentences depends on how certain they are about their response.

Moreover, these studies question the **uniform cost of the SIs computation** as illustrated in Figure 1 by H1. Instead, they suggest that the presence and magnitude of the delay effect depend on the proportion of literal and pragmatic responders in the tested sample. Considering the differences between consistent and inconsistent responders, the delay effect may arise in three different scenarios. Under H2, the delay effect would arise if **a large number of literal responders** provided fast *true* responses and slow *false* responses. An equal number of consistent pragmatic and literal responders hinders the delay effect. In scenario H3, the delay effect is driven by fast *true* responses from literal responders. The computation of SI is costly; therefore, the pragmatic responses are slower, and inconsistent responses indicate **the disambiguation of the sentence between the two readings**. In scenarios H2 and H3, many inconsistent responders might obscure the delay effect. Finally, under H4, the delay effect is due to **fast errors** made by inconsistent responders, for example, when they choose the literal response before finishing the SI computation. Under this hypothesis, the delay effect would arise when there are more inconsistent than consistent responders. In turn, when only consistent responders are present, the reversed delay would be present.

In summary, the delay effect might stem from different configurations of response patterns. To fully understand the source of this effect, individual differences in the choice of reading should be considered. Moreover, because the processing cost of SI computation varies across scalars, the effect of individual differences in the reading choice should also be examined across different scales.

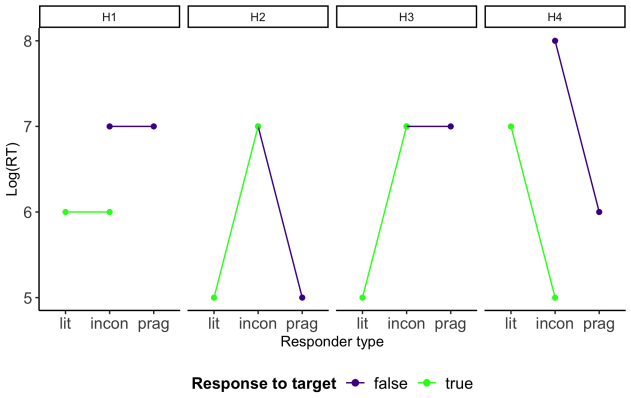

Figure 1: Four possible patterns that lead to the delay effect: H1 uniform cost of SI, H2 ratio between literal vs. pragmatic responders, H3 disambiguation of readings, H4 fast errors (lit-literal; prag-pragmatic; incon - inconsistent responders.

Present paper

This paper aims to test H1-H4 among different scales for which the delay effect was observed. To this end, we reanalyzed four datasets from four verification tasks testing different scales: three published by Van Tiel et al. (2016) and one by Van Tiel and Pankratz (2021) and made available at the Github and OSF repositories associated with the papers. To quantify individual differences in reading preferences, we applied the computational model proposed by Ramotowska, Marty, Van Maanen, and Sudo (2024) to the response data from the four experiments. This model allows for estimating the fine-grain measure of the by-subject probability of being a literal responder. To test the overall effect of reading preferences on the speed of computation of SI, we clustered participants into groups based on the model parameters. To determine between H1 and H2/H3, we tested the interaction effect between response type and reading preferences. To assess whether inconsistent responses are slower than consistent responses (H2 vs. H3 vs. H4), we tested whether the speed of *true/false* responses depends on the type of responder. Finally, we used the fine-grain measure of reading preferences as a predictor of SI verification speed, for which the

(reversed) delay effect was reported.

Methods

Experiments

Exp 1 We re-analyzed the data from the first and second picture-sentence verification online experiments reported in Van Tiel et al. (2019) (henceforth Exp 1). All experiments were in English and involved 7 scales: $< low, empty >, < scarce, absent >, < or, and >, < might, will >, < some, all >, < most, all >$, and $< try, succeed >$ two of which were classified by the authors as negative (scalars: low and scarce) and 5 positive[1]. In the first verification task experiment (henceforth NO-LOAD; N = 50), participants self-paced read a sentence and then saw a picture. They judge whether the sentence is a "good" or a "bad" picture description. In the second experiment, in addition to the verification task, participants had to perform a working memory task. They were randomly assigned to a low or a high working memory load condition (henceforth LOW-LOAD, HIGH-LOAD; N = 100).

Exp 2 Similarly to EXP 1, the experiment reported in Van Tiel and Pankratz (2021) (henceforth Exp 2; N = 50) was an online sentence-picture verification task in English. The experiment involved 16 adjectival scales classified by the author as positive ($< content, happy >, < fair, good >, < passable, good >, < ajar, open >, < chubby, fat >, < warm, hot >$) or negative ($< ripe, overripe >, < scarce, absent >, < sleepy, asleep >, < unlikely, impossible >, < youthful, young >, < breezy, windy >, < cool, cold >, < drizzly, rainy >, < low, empty >, < mediocre, bad >$). The procedure in this experiment was the same as in the NO-LOAD experiment.

In all experiments, responses and reaction times were recorded. All experiments involved 3 repetitions of the SI-triggering sentence and additional control sentences. We applied the same exclusion criteria for participants as in the original studies, resulting in sample sizes of: Exp1 48, 45, and 40; Exp2. 47.

Modeling individual differences

We applied the hierarchical beta-binomial Bayesian model with the latent group classification parameter (Ramotowska, Marty, et al., 2024). The model assumes two underlying groups of responders, literal and pragmatic, with different response distributions: $a_{lit} \sim Uniform(n/2, n)$ and $a_{prag} \sim Uniform(0, n/2)$ respectively (n is a total number of responses). The responses provided by a i-ty participant to j-ty SI come from the Beta distribution $\beta_{ji} \sim logit(Beta(a_{zji}, b_{zji}))$, where a_{zji} depends on the group classification parameter z_{ij} and $b_{zji} = n - a_{zji}$. The z_i-parameter indicates how likely each i-ty participant is a literal or pragmatic responder. By introducing a different number of classifications k, we can test the consistency of reading preference across scales (z_{ik}). z_{ik} comes form $Bernoulli(q_k)$ distribution, where q_k has a hyperprior Beta distribution $Beta(1,1)$.

We fit a series of models to the responses from Exp 1 with an increasing number of k-parameters $\{1, 2, ..., 7\}$[2]. For example, the model with $k = 1$ assumed that participants had the same reading preferences for each scale, while the model with $k = 7$ assumed that they had different preferences for each scale. Clustering and correlation analyzes of Van Tiel et al. (2019) suggested that participants had more consistent response preferences for scales *or*, *most*, *some*, and *might* than for other scales. We tested thus if the k parameter could be constrained among these scales.

Because increasing the number of parameters increases the complexity of a model, we applied the model comparison to establish which model achieves the best trade-off between complexity and fit. We used the Deviance Information Criterion (DIC Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) and the Watanabe-Akaike Information Criterion (WAIC Watanabe, 2010). The justified increasing complexity of the model would indicate large interindividual differences.

We used R JAGS (Plummer et al., 2003) to fit the computational model. No responses to SI-triggering sentences were excluded. The models were fit with the *jags* function, and the JAGSUI package was used to obtain WAIC values. Each fit involved 6 Markov chains with 10000 iterations. We applied 1000 burn-in iterations per chain and 1000 adaptive phase iterations.

Clustering and RT analyses

All analyses were conducted in R Studio (RStudio Team, 2020). We conducted four exploratory hierarchical clustering analyzes (R function HCLUST, using the ward.D2 method) on z_{ij}-parameters in each dataset. We used the R function NBCLUST (Charrad, Ghazzali, Boiteau, & Niknafs, 2014) to determine the appropriate cluster solution.

We preprocessed the data and applied thresholds for fast and slow reaction times (RT) as reported in the original papers. We fit the linear mixed-effects models using the LMERTEST package (Kuznetsova, Brockhoff, & Christensen, 2017). In the first analysis, all linear mixed-effects models included response type to SI-triggering sentence (true vs. false), cluster membership, and their interaction, and by-subject random intercept (and by-item random intercept in Exp 1) as predictors of the log-transformed RT in the SI-triggering condition in each tested dataset. In the second analysis, the log-transformed RT was the dependent variable and response type to the SI-triggering sentence (true=1 vs. reference level false=0), z_{ij}-parameter of the best-fitting models for that scale, and their interaction as predictors (by-subject Exp 1 and 2 and by-item random intercepts Exp 1). Because z_{ij}-parameter is a continuous predictor that indicates the probability of classification as a literal responder, we used a linear

[1] Van Tiel et al. (2019) also tested the scale $< may, must >$ in the first experiment but did not analyzed this data. We also did not include this scale in our analysis.

[2] The details about each model specification and the model comparison can be found on OSF.

transformation (z_{ij}-0.5) to interpret the response type effect with respect to inconsistent responses. The significant response effect would be in favor of H1 / H4 over H2 / H3. Finally, we tested the effect of the z_{ij}-parameter for *true/false* responses separately, to determine between H2-H4.

Results

Modeling results

All models converged with Rhats < 1.1. According to WAIC values, the best-fitting model for the NO-LOAD condition[3] had separate classification parameters for each scale. We selected the most complex model for further analysis based on WAIC values for all three working memory conditions and for Exp 2. Figure 2 shows the fit of the selected model in the NO-LOAD condition.

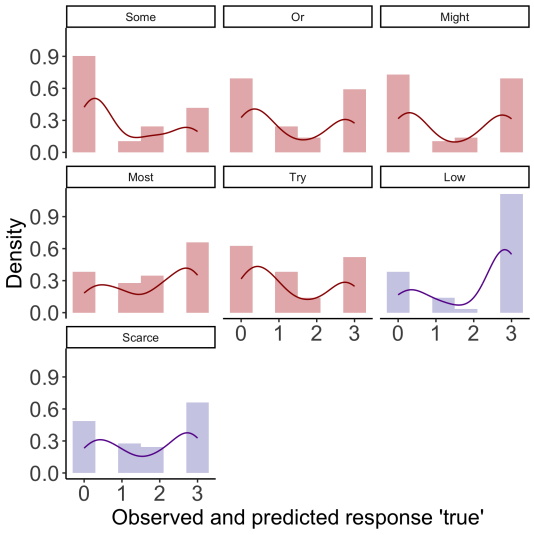

Figure 2: Fit of the most complex model in the NO-LOAD condition. The histograms show the observed data, and the lines show the model fits. Red color indicates the positive scales and purple color the negative scales.

Clustering of responders

The NBCLUST function indicated the two-cluster solution in all three working memory load conditions, as well as Exp 2. Visual inspection of the dendrogram confirmed that the two-cluster solution was accurate.

Exp 1 In the NO-LOAD condition, 25 participants were classified into the first cluster with a mean z_{ij}-parameter of 0.71 and 23 into the second cluster with a mean z_{ij}-parameter of 0.24. In the LOW-LOAD condition, 25 participants constituted the first cluster with a mean z_{ij}-parameter of 0.84 and 20 in the second cluster with a mean z_{ij}-parameter of 0.32. Finally, in the HIGH-LOAD condition, 28 participants were classified into the first cluster with a mean z_{ij}-parameter of 0.83 and 12 into the second cluster with a mean z_{ij}-parameter of 0.32. We call the first cluster a literal responder and the second one a pragmatic responder cluster.

For exploratory purposes, we plot the distributions of the z_{ij} parameters for each scale in the NO-LOAD condition (see Figure 3) with the cluster membership indicated. We observed that the individuals were inconsistent with their choice of response across the scales[4]. This complements the model comparison results and justifies the choice of the most complex model.

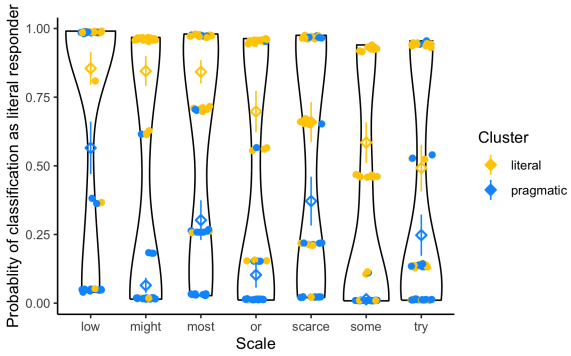

Figure 3: The distribution of z_{ij}-parameters in each scale in the NO-LOAD condition. The yellow color indicates participants clustered as literal responders, and blue as pragmatic responders. The diamonds show the mean z_{ij}-parameter per scale and cluster, and the bars indicate SE.

In the NO-LOAD condition, the RT analyzes revealed a significant effect of response type and an interaction between response type and cluster. The literal responders were faster to respond *true* than *false* (M = 7.01 vs 7.25; $\beta = -.18; t = -3.53; p = 0.005$), while the pragmatic responders were equally fast for both types of responses (M = 7.15 vs 7.27; $\beta = .08; t = 1.45; p = 0.15$).

Table 1: Summary of the RT analyses in Exp 1 and 2 (abbreviations: RESP - response effect; CLUST - Cluster effect; INTER - interaction effect; * < 0.05, ** p < 0.01, *** p < 0.001).

Exp	Intercept	RESP	CLUST	INTER
Exp 1 No	7.2***	-0.18**	-0.04	0.26***
Exp 1 Low	7.93***	-0.48***	-0.12	0.24**
Exp 1 High	7.72***	-0.11*	-0.11	0.4***
Exp2	7.09***	-0.14**	0.08	0.17*

The results of the LOW-LOAD condition showed a simi-

[3]Results of model comparison for LOW-LOAD and HIGH-LOAD conditions are available on OSF. The most complex models had the lowest WAIC values: LOW-LOAD 506 (DIC = 549.1) and HIGH-LOAD 459 (DIC 495.6).

[4]Similar inconsistencies were observed in two other working memory conditions.

lar pattern; the literal responders responded faster *true* than *false* (M = 7.59 vs 7.76; $\beta = -.24; t = -4.65; p < 0.001$), while the difference in response type was not significant for pragmatic responders (M = 7.78 vs 7.66; $\beta = -.006; t = -0.1; p = 0.92$). In the HIGH-LOAD condition, the literal responders were faster when they answered *true* than *false* (M = 7.61 vs 7.72; $\beta = -0.11; t = -2.21; p = 0.03$), and pragmatic responders had faster *false* response (M = 7.61 vs 7.88; $\beta = 0.29; t = 4.24; p < 0.001$).In general, the results NO/LOW-LOAD conditions are in line with H3, while those of the HIGH-LOAD condition are in line with H2.

Exp 2 The first cluster included 23 participants with a mean z_{ij}-parameter of 0.71, and the second cluster included 24 participants with a mean z_{ij}-parameter of 0.36. RT analysis revealed a significant effect of response type and an interaction between response type and cluster membership.[5] Subsequent analyses of literal cluster members revealed a significant effect of response type (faster *true* than *false* response $\beta = -0.13; t = -4.24; p < 0.001$) and lack of such an effect for pragmatic responders ($\beta = 0.005; t = 0.16; p = 0.87$) in line with H3.

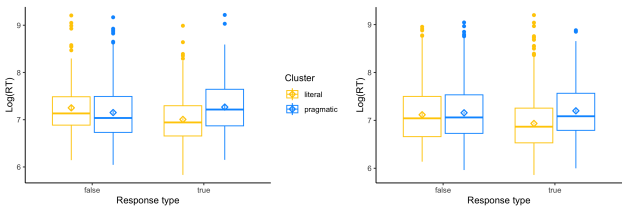

Figure 4: The log-transformed RTs in the NO-LOAD condition (left) and Exp 2 (right) as a function of response type (*true* vs. *false*). The yellow color indicates participants clustered as literal responders, and blue as pragmatic responders. The diamonds show the mean RT, and the bars indicate SE.

The delay effect analysis across scales

Next, we analyzed the source of the delay effect in individual scales in Exp 1: *or, might, most* and reversed delay effect for *scarce*; and in Exp 2, the delay effect for: *content, fair, passable, ajar, chubby, warm, youthful*.

Exp 1 Table 2 summarized the results for all LOAD conditions. The analysis in the NO-LOAD condition revealed an interaction effect for *or* and a marginal interaction for *most*. In the LOW-LOAD condition, we found significant interaction for *most*, and in the HIGH-LOAD condition for *or*. In NO-LOAD and LOW-LOAD conditions, we observed a significant effect of response type for *scarce*, and in the HIGH-LOAD condition, a significant effect of z_{ij}-parameter for *or*. Moreover, the z_{ij}-parameter effect was present for the *true* responses in the case of *or* ($\beta = -0.51; t = -2.39; p = 0.02$; $\beta = -1.06; t = -2.44; p = 0.02$) and *scarce* ($\beta = -.52; t = -2.12; p = 0.04$; $\beta = -.7; t = -2.2; p = 0.03$) in NO-LOAD and LOW-LOAD conditions, respectively; and in the case of *might* ($\beta = -1.60; t = -2.7; p = 0.01$) in the HIGH-LOAD condition (and marginal effect for *or* $\beta = -0.78; t = -1.79; p = 0.08$). Thus, although the results are not fully consistent across LOAD conditions, for *or* the significant interaction, the lack of the response type effect, and the z_{ij}-parameter effect for *true* responses suggest that these scalars follow the pattern under H3; however, under HIGH-LOAD, *or* patterns more as in H2 (due to z_{ij}-parameter also for *false* responses). The pattern for *most* is similar, however, less clear due to the lack of a significant effect of z_{ij}-parameter in the *true* responses. In contrast, for *might*, the lack of interaction and z_{ij}-parameter effect suggests that it patterns with H1 predictions; however, under HIGH-LOAD it patterns more with H3. In contrast, for *scarce* we observed the lack of interaction, the response effect, and the z_{ij}-parameter effect for *true responses*. Given that the delay effect for *scarce* was reversed, the results suggest that it patterns with reversed H4.

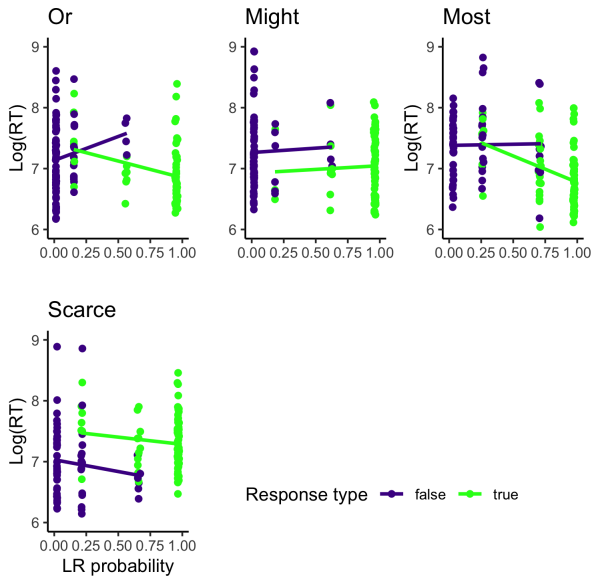

Figure 5: The log-transformed RTs to SI-triggering sentence in the NO-LOAD condition as a function of z_{ij}-parameter. The green color indicates response *true* and purple *false*. The lines show the effect of the z_{ij}-parameter on each response type. LR stands for literal responder.

Exp 2 The analyses revealed that the response preferences affected the reaction times for scalars *chubby* ($\beta = -2.2; t = -3.79; p < 0.001$) (for *ajar*, the p-value for interaction effect was 0.0501). Moreover, the effect of z_{ij}-parameter for *chubby* ($\beta = .99; t = 2.8; p = 0.01$). In the case of *true* responses, the z_{ij}-parameter was a significant predictor for *chubby* ($\beta = -1.1; t = -2.99; p = 0.004$) and *warm* ($\beta = -1.5; t = -2.2; p = 0.03$). Thus, *chubby* patterned with H2, *content, fair, passable, youthful* with H1, *ajar, warm* (less clearly) with H3.

[5] This model also included the by-subject random slope for response type.

Table 2: Summary of the RT analyses all LOAD conditions (abbreviations: RESP - response effect; INTER - interaction effect; . < 0.1, * < 0.05, ** p < 0.01, *** p < 0.001).

Load	Scale	RESP	z_{ij}	INTER
No	Scarce	0.63***	-0.35	-0.15
No	Might	-0.34	0.09	-0.0007
No	Or	-0.4.	0.79.	-1.3*
No	Most	-0.2.	-0.03	-0.75
Low	Scarce	0.31*	0.3	-0.9.
Low	Might	-0.46	-0.06	0.44
Low	Or	0.16	-0.01	-0.98
Low	Most	-0.2	0.6	-1.1*
High	Scarce	0.20.	0.48	-0.43
High	Might	0.52	0.07	-1.6.
High	Or	-0.01	0.88*	-1.63**
High	Most	-0.16	-0.11	-0.16

Discussion

In this study, we re-analyzed the responses and reaction times from four verification task experiments. Three experiments included 7 scalars (5 positive and 2 negative, Van Tiel et al., 2019), and one dataset included 16 adjectival scalars (6 positive and 10 negative Van Tiel & Pankratz, 2021). We applied a computational model (Ramotowska, Marty, et al., 2024) that allowed us to quantify individual differences in reading choice by probabilistically classifying participants as literal or pragmatic responders for each scale. The results of the modeling revealed substantial intra-individual variability in the selection of the reading. In particular, we found that the most complex model with by-scale classification parameters was justified.

Next, we tested whether individual differences in response preferences play a role in the speed of verification. Across datasets, we found that the literal responders answer faster *true* than *false* as predicted by H3, and, in one case, also that pragmatic responders respond faster *false* than *true* as predicted by H2. However, analyzis of individual scales revealed that some scales also patterned in line with H1 and H4.

Scalar diversity vs. individual differences

Contrary to previous findings (Ramotowska, Marty, et al., 2024), our results did not show that SI-triggering sentence polarity modulates participants' reading choices. Ramotowska, Marty, et al. (2024) found more uncertain participants for negative SI-triggering sentences and more literal responders for the positive SI-triggering sentence. One reason for the differences between our and Ramotowska, Marty, et al. (2024) results could be that in their case, the polarity of the SI-triggering sentence was given explicitly with negation (*some not; not all*), while in the case of Van Tiel and Pankratz (2021) and Van Tiel et al. (2019) studies, the polarity of the scalar was implicit.

Our results suggest that individual differences should be considered when seeking an explanation of scalar diversity. For example, future studies could investigate whether individuals are sensitive to different linguistic factors that trigger SI. In addition, studies should focus on developing models that explicitly link scalar diversity with individual differences. The Item Response Models, which have been successfully applied to investigate individual differences in categorization (Verheyen, Hampton, & Storms, 2010) and quantifier meaning representations (Ramotowska, Haaf, Van Maanen, & Szymanik, 2024), may be a promising direction.

The delay effect

Our results clearly show that the responder preferences modulate the speed of SI processing; however, the direction of this modulation might vary across scales. The general pattern of results supports H3, locating a part of the delay effect in the disambiguation between literal and pragmatic meanings. However, when zooming into individual scales, we observed that in some cases the SI processing cost is independent of the responder type (H1), and in one case, the pragmatic responders tend to make fast errors (for the reversed delay effect of *scarce* in Exp 1).

More research is needed to replicate these results and confirm that the variation between scales is not an artifact of a small number of inconsistent responses. More repetition of the SI-triggering items would be desired to obtain more reliable results. It is well known that the reaction time distribution is skewed with a long tail of slow reaction times (Ratcliff, 1993). Because of that, slow reaction times, typically uncertain responses, are rare. A larger number of repetitions would allow us to capture a whole distribution of reaction times. With more repetitions of the same item, we could also obtain slower reaction times and wider ranges of responders with less certain preferences toward literal or pragmatic reading.

Our findings suggest that the delay effect should be reconsidered in light of individual differences. In particular, the reaction times and choices should be modeled jointly. Evidence accumulation models, such as the Diffusion Decision Model (DDM, Ratcliff & McKoon, 2008), make this possible. The DDM's parameters allow testing several additional predictions, e.g., faster responses for preferred reading are due to initial response bias, vs. the decision process is faster for one of the readings. The model may also explain the shift to literal reading under speed pressure (Bott & Noveck, 2004; Ratcliff & McKoon, 2008).

Conclusion

In this study, we investigated four explanations of the delay effect across different scales. The results suggest that the delay effect might have a different source across different scales.

Acknowledgments

The author was supported by the NWO OC project Nothing is Logical (grant no 406.21.CTW.023). Many thanks to Paul Marty for the discussion about this project.

References

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

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, *51*(3), 437–457.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36.

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology*, *64*(12), 2352–2367.

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

Feeney, A., & Bonnefon, J.-F. (2013). Politeness and honesty contribute additively to the interpretation of scalar expressions. *Journal of Language and Social Psychology*, *32*(2), 181–190.

Gotzner, N., Solt, S., & Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in psychology*, *9*, 1659.

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

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. R. (1989). A natural history of negation.

Khorsheed, A., & Gotzner, N. (2023). A closer look at the sources of variability in scalar implicature derivation: a review. *Frontiers in Communication*, *8*, 1187970.

Khorsheed, A., Price, J., & van Tiel, B. (2022). Sources of cognitive cost in scalar implicature processing: A review. *Frontiers in Communication*, *7*, 990044.

Kursat, L., & Degen, J. (2020). Probability and processing speed of scalar inferences is context-dependent. In *Annual meeting of the cognitive science society* (p. 1236-1242).

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Plummer, M., et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, pp. 1–10).

Ramotowska, S., Haaf, J., Van Maanen, L., & Szymanik, J. (2024). Most quantifiers have many meanings. *Psychonomic Bulletin & Review*, 1–12.

Ramotowska, S., Marty, P., Van Maanen, L., & Sudo, Y. (2024). Some but not all speakers sometimes but not always derive scalar implicatures. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological bulletin*, *114*(3), 510.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

Ronai, E., & Xiang, M. (2022). Three factors in explaining scalar diversity. In *Proceedings of sinn und bedeutung* (Vol. 26, pp. 716–733).

RStudio Team. (2020). Rstudio: Integrated development environment for r [Computer software manual]. Boston, MA.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Harvard University Press Cambridge, MA.

Sperber, D., & Wilson, D. (1987). Précis of relevance: Communication and cognition. *Behavioral and brain sciences*, *10*(4), 697–710.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583–639.

Van Tiel, B., & Pankratz, E. (2021). Adjectival polarity and the processing of scalar inferences. *Glossa: a journal of general linguistics*, *6*(1).

Van Tiel, B., Pankratz, E., & Sun, C. (2019). Scales and scalarity: Processing scalar inferences. *Journal of Memory and Language*, *105*, 93–107.

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of semantics*, *33*(1), 137–175.

Verheyen, S., Hampton, J. A., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the rasch model. *Acta psychologica*, *135*(2), 216–225.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(116), 3571–3594.