Representational complexity and pragmatics cause the monotonicity effect

Fabian Schlotterbeck* (fabian.schlotterbeck@uni-tuebingen.de)
University of Tübingen

Sonia Ramotowska* (s.ramotowska@uva.nl)
University of Amsterdam

Leendert van Maanen (lvmaanen@gmail.com)
Utrecht University

Jakub Szymanik (jakub.szymanik@gmail.com)
University of Amsterdam
*Co-first authors.

Abstract

Psycholinguistic studies have repeatedly demonstrated that downward entailing (DE) quantifiers are more difficult to process than upward entailing (UE) ones. We contribute to the current debate on cognitive processes causing the monotonicity effect by testing predictions about the underlying processes derived from two competing theoretical proposals: two-step and pragmatic processing models. We model reaction times and accuracy from two verification experiments (a sentence-picture and a purely linguistic verification task), using the diffusion decision model (DDM). In both experiments, verification of UE quantifier *more than half* was compared to verification of DE quantifier *fewer than half*. Our analyses revealed the same pattern of results across tasks: Both non-decision times and drift rates, two of the free model parameters of the DDM, were affected by the monotonicity manipulation. Thus, our modeling results support both two-step (prediction: non-decision time is affected) and pragmatic processing models (prediction: drift rate is affected).

Keywords: monotonicity; quantifiers; semantic representations; pragmatics; diffusion decision model

Background and goals

Psycholinguistic studies have repeatedly demonstrated that downward entailing (DE) quantifiers are more difficult to process than upward entailing (UE) ones. While this monotonicity effect was found in a range of different cognitive tasks, such as reading and reasoning, it shows up most reliably in verification tasks (e.g. Clark, 1976; Deschamps, Agmon, Loewenstein, & Grodzinsky, 2015; Just & Carpenter, 1971; Szymanik & Zajenkowski, 2013). Although the empirical phenomenon itself is well-documented, it is a matter of current debate which cognitive processes cause the monotonicity effect (e.g Agmon, Loewenstein, & Grodzinsky, 2019; Nieuwland, 2016; Schlotterbeck, 2017). Our main aim is to contribute to this debate by testing predictions about the underlying processes derived from two competing theoretical proposals: two-step and pragmatic processing models. To this end, we model data from two verification experiments, in particular, reaction times (RT) and accuracy, using a well-established model of decision making from mathematical psychology, namely the diffusion decision model (DDM, see e.g. Ratcliff, 1978; Ratcliff & McKoon, 2008).

Competing theoretical proposals

Various explanations of the monotonicity effect have been proposed in the literature. We distinguish between two broad classes here. Explanations in the first class (two-step processing models) are based on an additional processing step in the verification of DE vs. UE quantifiers. The earliest two-step models (e.g. Just & Carpenter, 1971) were derived from the basic hypothesis that contexts and sentence meanings are both mentally encoded in a symbolic propositional format that can then be compared to each other symbol by symbol in a verification task. The monotonicity effect is explained by the assumption of a negation symbol present in the encoding of DE but not UE quantifiers, which corresponds to an extra step in the verification process. More recent alternatives make somewhat different assumptions, e.g., about the processing of negation (cf. Kaup, Zwaan, & Lüdtke, 2007) or the involved meaning representations (e.g. Deschamps et al., 2015; Schlotterbeck, 2017), but share the assumption of an additional computational step.

A radically different view is taken by accounts that rely on a pragmatic processing model (e.g. Degen & Tanenhaus, 2019), which assumes that contextual fit or pragmatic felicity is a major determinant of processing difficulty. Under this view, DE quantifiers cause processing difficulties because they are systematically dispreferred to suitable UE alternatives in various contexts (cf. Nieuwland, 2016; and also Nieuwland & Kuperberg, 2008; for an analogous view on the processing of negation) due to violation of pragmatic principles (e.g. avoidance of infrequent words or uninformative statements, cf. Grice, 1975). In order to draw an explicit connection between pragmatic considerations of this kind and data from verification tasks, verification is often thought of as production: Participants in a verification task, in fact, judge whether they would utter the sentence to describe the context (e.g. Degen & Goodman, 2014; Waldon & Degen, 2020). Recent Bayesian models of rational speaker behavior (e.g. Frank & Goodman, 2012) allow us to formalize the effects of factors such as word frequencies or informativity on speakers' production probabilities. In this way, the monotonicity effect can be explained without assuming an additional processing step (cf. Nordmeyer & Frank, 2014, for a related proposal).

Main ingredients of the DDM

In the DDM, decision processes, such as true/false judgments, are described as the accumulation of a noisy signal over time

until a decision boundary is reached and a response is initiated. One main strength of the DMM is that it concurrently models both accuracies and entire RT distributions. Moreover, its free model parameters correspond to distinct components of the underlying cognitive processes. The estimation of these parameters, therefore, allows inferences about the processing components involved in the experimental task. The DDM parameters represent independent processing components, meaning that each parameter explains different RT and accuracy effects. In this way, the DDM allows to model independent sources of variation between conditions. For the present purpose, the most important parameters are *drift rate* (v) and *non-decision time* (T_{er}). Drift rate determines how much information is accumulated per time unit and non-decision time measures RT components that are not themselves part of the decision process, e.g. processes related to the stimulus encoding or execution of a motor response. In addition, the standard DDM model has also a parameter, a, which specifies the separation between the two decision boundaries; a parameter, z, which determines where between the two boundaries decision processes will start, and variability parameters (s_z, s_{Ter} and s_v), which allow for trial-to-trial variability of starting point, non-decision time and drift rate, respectively. In this paper, we focus on drift rate and non-decision time parameters, which are closely related to the cognitive processes of interest. The a parameter is usually used to model speed-accuracy trade-off (fast responses, more errors vs. slow responses, less errors) and z parameter to model response bias (starting points can be closer to one of the boundaries) (e.g Mulder, van Maanen, & Forstmann, 2014; Ratcliff & McKoon, 2008). These two parameters do not explain the typical patterns of RT and accuracy in verification of DE and UE quantifiers.

The DDM is a theoretically well-founded model (e.g. Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006) that has been applied successfully to a large variety of decision tasks (for review, see e.g. Mulder et al., 2014; Ratcliff & McKoon, 2008). For example, a good model fit was observed in previous studies that applied the DDM to RT and accuracy collected in number comparison tasks (e.g. Dehaene, 2007; Ratcliff & McKoon, 2018). As there are close similarities between number comparison and verification of proportional quantifiers, the DDM is, therefore, a natural choice to model the latter task as well. These previous studies found that drift rate is monotonically related to numerical distance, with larger drift rates for numerosities that are further apart from each other. In comparison tasks that involved the precise comparison of numerals, a step-like relationship was observed. For approximate numerosities, drift rates were in a linear relationship with the logarithm of the ratio (log ratio) of the two involved numerosities. These findings are consistent with current theories on the representation and processing of precise and approximate number (e.g. Feigenson, Dehaene, & Spelke, 2004) and they are also relevant for the comparison between the experiments reported below.

Link to theoretical proposals

One way to link two-step processing models to components of the DDM is to assume that monotonicity affects non-decision time in verification tasks because the truth evaluation of DE quantifiers involves an extra step in addition to the actual verification step (see Donkin, Heathcote, Brown, & Andrews, 2009, for related discussion and empirical data from lexical decision). For example, we could think of the verification of DE quantifiers as falsification of a suitable UE counterpart followed by a subsequent, time-consuming step of truth-value reversal. However, this extra step does not change the complexity of the underlying, non-negated representation and, therefore, should not affect drift rate.

By contrast, pragmatic models hold that DE quantifiers take longer to evaluate because they are generally dispreferred as descriptions of the presented contexts. Taking into account what evidence accumulation models like the DDM have revealed about processes in closely related domains, e.g. lexical selection in picture naming tasks (e.g. Anders, Riès, van Maanen, & Alario, 2015; Anders, van Maanen, & Alario, 2019), pragmatic models let us expect that monotonicity affects drift rates: Slower accumulation is expected for DE vs. UE quantifiers. This assumption is further motivated by theoretical considerations (e.g. Bitzer, Park, Blankenburg, & Kiebel, 2014; Bogacz et al., 2006) that allow us to relate parameters of the DDM (drift rate, specifically) to Bayesian pragmatic models predicting utterance production probabilities from factors such as word frequencies or informativity.

Methods

We conducted two web-based experiments, in which we compared the verification of UE quantifier *more than half* (*mth*) to DE quantifier *fewer than half* (*fth*). We decided to use two different paradigms - one visual (i.e. sentence-picture) and one purely linguistic (i.e. sentence-sentence) verification task. By comparing these two paradigms we were able to not only test the robustness of the effects but also their linguistic relevance. In particular, the sentence-picture experiment involves both linguistic and visual processing. By showing that similar effects occur in both setups we provide an extra evidence for the linguistic character of the effects. Additionally, while the purely linguistic experiment may rely more on the precise comparison of involved numerosities the visual experiment is most likely relying on approximate numbers (see Szymanik, 2016, for discussion). Hence, our results also show that the monotonicity effect is not restricted to only approximate or precise processing of numerosities (cf. Dehaene, 2007). In both experiments, we collected the participants' responses and RT.

Participants

For the linguistic experiment, we collected data from 90 participants via Amazon Mechanical Turk (compensation \$4). The final sample (see "exclusion criteria") included 72 English native speakers (24 female, mean age 35 yr; $sd = 11$;

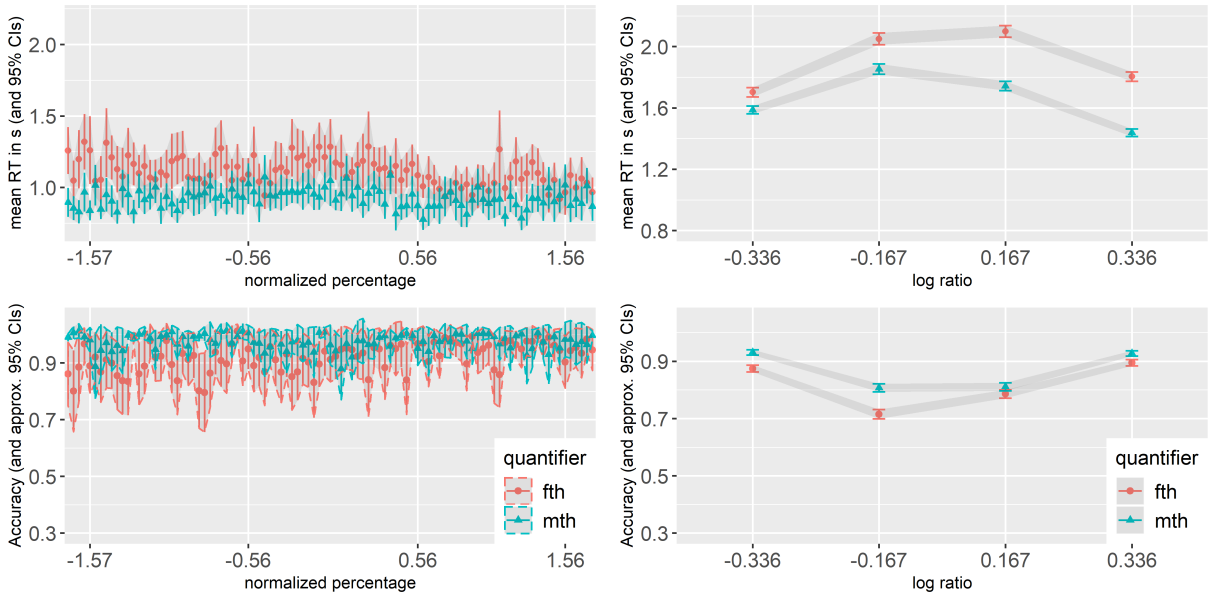

Figure 1: Descriptive results. Left: linguistic task; right: visual task.

Table 1: Results of regression analyses. MON: effect of monotonicity; MON × TV: truth value × monotonicity interaction.

	RT						Accuracy					
	linguistic task			visual task			linguistic task			visual task		
	β	t	p	β	t	p	β	z	p	β	z	p
MON	136	5.55	$< .001$	231	15.32	$< .001$	1.01	5.09	$< .001$	.19	3.28	.001
MON × TV	107	5.50	$< .001$	26	2.51	.012	.55	2.11	.035	.32	4.54	$< .001$
true conditions only												
MON	224	7.66	$< .001$	307	14.47	$< .001$	1.19	3.51	$< .001$	.55	9.24	$< .001$
false conditions only												
MON	151	5.71	$< .001$	246	11.44	$< .001$	1.32	3.27	.001	.14	2.24	.025

range: $22 - 59$). Participants of the visual experiment were recruited via `prolific.co` (compensation £7.5). Data from 96 English native speakers was collected in total and after exclusion the final sample consisted of 56 participants (49 female; mean age 36 yr; $sd = 13$; range: $18 - 69$).

Design, materials & procedures

Linguistic experiment (N=72, 50 trials per quantifier): Participants were presented with two sentences: a simple quantified sentence of the form "Q of the As are B", where "Q" was either *mth* or *fth* and "As" and "B" were pseudowords (e.g *glerbs* and *fizzda*) generated from English nouns and adjectives (Keuleers & Brysbaert, 2010); and a sentence of the form "X% of the As are B", where "X%" was a precise percentage between $1 - 99\%$, excluding 50%. The original 6-letter nouns and adjectives were controlled for frequency (Zipf value: 4.06; van Heuven, Mandera, Keuleers, & Brysbaert, 2014). The generated pseudowords were assessed by an English native speaker. In each trial participants saw a different pair of pseudowords. We also included filler trials with the quantifiers *most*, *many* and *few*. For *mth* and *fth* percentages were counterbalanced between percentages above and below 50%. Participants read the first sentence self-paced and their task was to decide if the first sentence is true given the information from the second. They responded by pressing one of two response keys on their keyboard. The experiment started with a short training block consisting of 8 trials with quantifiers that were not presented in the main experiment (i.e., *some*, *all*, *none*).

Visual experiment (N=56, 240 trials per quantifier): Participants first read a sentence like, e.g., *more than half of the dots are blue* self-paced and then evaluated it against a visual display showing blue and orange dots. Participants were instructed to judge as fast as possible whether the sentence is an appropriate description of the depicted quantitative relations. They provided their response by pressing one of two keys on their keyboard. A factorial within-participants design was used in which the two factors MONOTONICITY (2 levels: *mth* vs. *fth*) and RATIO of the colored dots (4 levels: 28:20, 26:22, 22:26 and 20:28) were crossed, yielding eight conditions. Each participant saw 60 trials in each con-

dition, amounting to a total of 480 trials. 480 pictures were generated by drawing colored dots at random positions in the two halves of a gray 512 px × 256 px background. The dots had a mean radius of 5.5 px (drawn from a normal distribution with $sd = 1$ and then clipped to the range $[1, 10]$). Which color was presented on which side of the picture was counterbalanced between items. Participants saw the same set of 60 pictures in the same conditions. In half of the items, the target color was blue, in the other half it was orange. Materials were presented in random order and distributed across four blocks. Each block consisted of roughly 120 trials, but the precise lengths of the four blocks were randomly chosen for each participant. In between blocks, there were self-paced breaks that participants initiated by pressing a button that they did not use otherwise. We recorded which button was pressed and thereby used the breaks as 'catch trials'. At the beginning of the experiment, there was a short practice session consisting of eight trials that were similar to the experimental trials but contained different quantifiers. In total, the visual experiment took participants about 40 min on average, roughly twice as long as the linguistic experiment. In both experiments, participants were randomly assigned to one of two possible response key mappings.

Exclusion criteria

Since data were collected over the web, we applied rather strict exclusion criteria in order to ensure high quality of the final data sets. These criteria were specified in advance and were based on the specifics of the two experiments (for discussion of data exclusion in the context of web-based experiments, see Kochari, 2019). In the linguistic experiment we excluded participants if they had more than 50% responses below 300 ms (fast guesses) or did not have increasing probability of saying 'true' ('false' for DE quantifiers) with increasing percentage (monotonicity violation). In addition, we excluded one more participant, who participated in a very similar study before. All together we excluded 18 participants.

In the visual experiment, the following criteria resulted in the exclusion of 40 participants. Participants were excluded if they had extraordinarily long reading times or RT (i.e. several minutes) in some trials; if they had more than five RT above 15 s or more than five reading times above 25 s; or if in more than one condition accuracy was not significantly above chance. In addition, we checked for participants that had many fast guesses or missed more than one of three catch trials (see procedure). All of the latter had, however, already been excluded by one of the other criteria.

In the linguistic task, we also excluded trials with RT faster than 300 ms or longer than mean+2*SD (calculated for true and false responses separately). In the visual task, we excluded trials with reading times or RT shorter than 200 ms or longer than mean+3.5*SD (calculated per participant and condition).

Regression analyses and modeling strategy

First, the data were analyzed using mixed effects regression models that mainly tested for two known effects: the monotonicity effect and the interaction between monotonicity and truth value (e.g. Just & Carpenter, 1971). To this end, independent variables were recoded in the following way. The analysis of the linguistic task included the absolute value of the normalized percentage (z-scored percentage with 50% as zero) as a numerical predictor and the analysis of the visual task included the absolute value of the logarithm of the ratio of the two presented numerosities in each trial (ABSOLUTE LOG RATIO) as a factor (levels: .167 vs. .336). In addition, analyses of both tasks included the factors MONOTONICITY (levels: *fth* vs. *mth*) and TRUTH VALUE (levels: *true* vs. *false*). Conditions with *mth* were coded as *true* if normalized percentage or log ratio was positive and as *false* if they were negative. For *fth*, TRUTH VALUE was coded the opposite way.

Afterwards, the DDM was applied to test the above predictions. We fit the DDM to data from the two experiments separately. To this end, we used the R package `rtdists` and performed maximum likelihood estimation of DDM parameters using particle swarm optimization. We estimated non-decision time (T_{er}), starting point (z), boundary separation (a) and drift rate (v). All variability parameters were set to 0. We assumed that log-ratio and normalized percentage are monotonically related to drift rates and specified this relation using the following generalized logistic regression function, where: V_l is a lower asymptote; V_u is an upper asymptote; s is a growth rate; p_0 is a midpoint; and p is normalized percentage or log-ratio.

$$v(p) = V_1 + \frac{V_1 - V_u}{1 + e^{-s(p-p_0)}}$$

Results

Mean RT and accuracies are shown in Figure 1. Below we report the results of the regression and DDM analyses. [1]

Regression Analyses

The main results of the regression analyses are given in Table 1. The MONOTONICITY effect as well as the MONOTONICITY×TRUTH VALUE interaction were replicated in RT and accuracy in both experiments. Mean RT were faster and accuracy was higher for *mth* than for *fth* (LINGUISTIC: 926 ms vs. 1110 ms and 97.7% vs. 92.3%; VISUAL: 1655 ms vs. 1913 ms and 86.9% vs. 81.8%). Moreover, these effects were more pronounced in the false than in the true conditions (LINGUISTIC: true: 233 ms and 7.7% difference; false: 125 ms and 3% difference; VISUAL: true: 289 ms and 7.5% difference; false: 231 ms and 2.9% difference). To test for effects of MONOTONICITY independently of TRUTH VALUE, we conducted separate analyses for the true and false conditions. The MONOTONICITY effect was significant in all cases.

[1] The data and analysis scripts of both experiments are made available on `https://osf.io/4d69v`

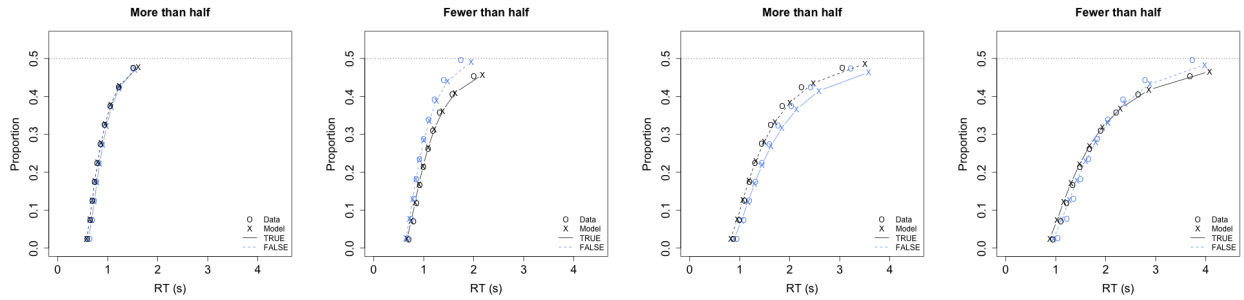

Figure 2: Defective CDF plots (Ratcliff, 1979) showing average model fit (first two: linguistic task; second two: visual task).

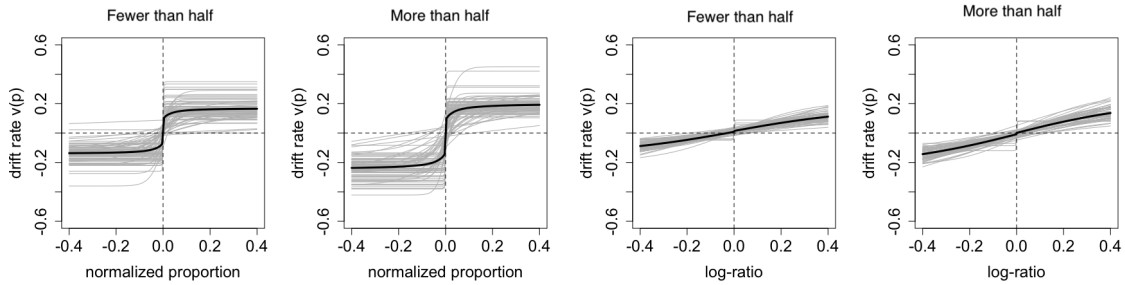

Figure 3: Relation between drift rates and numerical information. Left: linguistic task; right: visual task. Gray lines correspond to individual participants; black lines are based on mean parameter estimates.

DDM analyses

First, we fitted the DDM to the linguistic data and used model comparisons (based on BIC Schwarz, 1978) to determine which parameters differed between quantifiers (see Table 2). We predicted that both quantifiers should have a 50% midpoint (p_0 parameter) and growth rate (s parameter), because the truth conditions for both quantifiers are unambiguously specified. Based on the patterns of RT and accuracy for both quantifiers, we did not find evidence for a speed-accuracy trade-off, typically modelled by the a parameter (Mulder et al., 2014; Ratcliff & McKoon, 2008). Therefore, we also constrained a to be the same for both quantifiers. We additionally tested that the constrained parameters did not differ between quantifiers s ($t(71) = .42; p = .68$), p_0 ($t(71) = -.96; p = .34$) and a ($t(71) = -1.45; p = .15$). The final model was the best model for 66 participants out of 72. Then, we applied the same model to the visual data. We verified that the model fit was good by examining participants individually. The overall model fit is shown in Figure 2.

Table 2: Summary of model constraining procedure

Model number	1	2	3	4
Constrained parameters	–	s	s, p_0	s, p_0, a
Number of free parameters	14	13	11	10
Model was best for:	0	1	5	66

In line with previous results (Dehaene, 2007), a comparison revealed that decision processes differed between the two tasks: Drift rate increased gradually with log-ratio in the visual task, whereas a step-like relation was found in the linguistic task (see Figure 3). Apart from this difference, we found consistent results across the two tasks. In both tasks, non-decision times were longer for *fth* than *mth* (LINGUISTIC: $t(71) = 5.53; p < .001$; VISUAL: $t(55) = 5.74; p < .001$). The mean difference between *fth* and *mth* was 34 ms in the linguistic and 43 ms in the visual task.

To test for differences in drift rates, we calculated distances between the asymptotes (V_u - V_l) of the logistic regression function. We found that the mean distances between asymptotes were larger for *mth* (LINGUISTIC: .46; VISUAL: .64) than for *fth* (LINGUISTIC: .31; VISUAL: .46). This means that drift rates were higher for *mth* than for *fth* (LINGUISTIC: $t(71) = 9.10; p < .001$; VISUAL: $t(55) = 8.46; p < .001$).

Moreover, we also tested for differences in relative starting points. In the linguistic task, we found a *yes*-bias for *mth* (the starting point was closer to the upper decision boundary) compared to *fth* (.56 vs. .49; $t(71) = 5.56; p < .001$). In the visual task, both quantifiers exhibited a *yes*-bias (.54 vs. .55; $t(55) = -.96, p = .34$).

Because model 4 was the best model for only 66 out of 72 participants, we tested additionally if the variation between participants in best model fit has an effect on our results. To test this we computed Bayesian model averaged (BMA) pa-

rameters. The BMA method takes into account parameters from all fitted models and computes weighted average parameters according to the models' BIC values (Wagenmakers & Farrell, 2004). The BIC weight w for model i is defined by the following equation, where $\Delta_i(BIC) = BIC_i - min(BIC)$.

$$w_i(BIC) = \frac{exp\{\frac{-1}{2}\Delta_i(BIC)\}}{\sum_{k=1}^{K} exp\{\frac{-1}{2}\Delta_k(BIC)\}}$$

We tested the difference between DE and UE quantifiers in non-decision time and drift rate parameters. We found the expected difference in non-decision time ($t(71) = 5.63; p < .001$), and drift rate ($t(71) = 9.50; p < .001$). These findings indicate that the variation between participants was negligible.

Discussion

We applied the DDM to data from two web-based verification experiments in order to test predictions derived from theoretical accounts of the monotonicity effect. From two-step accounts, we derived the prediction that non-decision time would be affected, and from pragmatic processing models, we derived the prediction that drift rate would be affected.

The monotonicity effect was replicated in both experiments, and our modeling results are entirely consistent across both experiments: we found that the monotonicity manipulation affected both parameters, drift rates and non-decision times, in the expected direction. Therefore, our results support both hypotheses and indicate two potential sources of the monotonicity effect that map onto different DDM parameters. Moreover, they show that the monotonicity effect and its cognitive correlates are robust across various linguistic tasks, strongly suggesting that they are inherent in language processing. We acknowledge that an unambiguous mapping from effects in non-decision times and drift rates to representational complexity and pragmatic processes, respectively, can be challenged. Nevertheless, our modeling results render accounts that explain effects on only one of the two parameters implausible, or at least incomplete.

Recently, Agmon et al. (2019) arrived at similar conclusions analyzing mean RT. They compared verification of quantifiers, e.g. *mth* vs. *fth*, to the verification of expressions containing positive vs. negative adjectives, e.g. *a large* vs. *a small proportion*. Like *fth*, *a small proportion* is also negative, but it is not DE. Across a range of comparable expressions, they found larger RT differences between pairs that differ along both of these dimensions, than between expressions that differ only in negativity. They argued that both negativity and downward monotonicity are sources of increased processing difficulty. One way to explain these findings in our present terms and also to explain the two sources of processing difficulty we observed in estimated DDM parameters would be to assume that negativity affects pragmatics. In contrast, only DE expressions involve an extra processing step. While the relevant theoretical distinctions are, in fact, more subtle than what we can cover here (see also Bott, Schlotterbeck, & Klein, 2019, for discussion), the empirical question how our modeling approach relates to these findings is interesting in its own right. We plan to address this question in ongoing efforts.

Another well-documented effect - the interaction between monotoncity and truth value - was also replicated in our experiments. Classical explanations of this effect are based on verification procedures (Barwise & Cooper, 1981; Szymanik & Zajenkowski, 2013; Deschamps et al., 2015). While the observed differences in mean RT, as well as our regression analyses, are consistent with previous findings, our modeling results are unexpected under those accounts: What our results indicate is a tendency to answer "yes, true" to *mth* in the visual and linguistic task. To obtain a better understanding of how response biases are related to the interaction between monotonicity and truth value, a comparison to the processing of negation may be instructive, where a similar interaction is often observed (Just & Carpenter, 1971).

Beside the mentioned similarities, we also found differences between the two tasks. As reflected in higher RT and lower accuracy, the visual task was the more difficult among the two. Moreover, the signature of the decision processes also differed between tasks (see Figure 3). These findings are consistent with existing studies (Dehaene, 2007) that applied the DDM to number comparison tasks involving either approximate (dot pictures) or precise numerosities (numerals). The fact that the present analyses replicate these results indicates that our method is sensitive enough to detect qualitative differences between tasks. Thus, the consistent results on monotonicity receive indirect validation.

Finally, our results demonstrate that decision models, like the DDM, are applicable to data collected over the web. We will take a closer look at this by comparison of our results to a replication in the lab.

Acknowledgements

We thank three anonymous reviewers, Rolf Ulrich and Milica Denić for helpful comments and discussions. FS has received funding from a postdoc fellowship of the German Academic Exchange Service (DAAD) under the programme number 57468851. SR and JS have received funding form the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement n. STG 716230 CoSaQ.

References

Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2019). Measuring the cognitive cost of downward monotonicity by controlling for negative polarity. *Glossa: A Journal of General Linguistics*, *4*(1), 36.

Anders, R., Riès, S., van Maanen, L., & Alario, F.-X. (2015). Evidence accumulation as a model for lexical selection. *Cognitive Psychology*, *82*, 57–73.

Anders, R., van Maanen, L., & Alario, F.-X. (2019). Multifactor analysis in language production: Sequential sam-

pling models mimic and extend regression results. *Cognitive Neuropsychology*, *36*(5-6), 234–264.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*, 159–219.

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. (2014). Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. *Frontiers in Human Neuroscience*, *8*, 102.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Bott, O., Schlotterbeck, F., & Klein, U. (2019). Empty-set effects in quantifier interpretation. *Journal of Semantics*, *36*(1), 99–163.

Clark, H. H. (1976). *Semantics and Comprehension*. Mouton.

Degen, J., & Goodman, N. D. (2014). Lost your marbles? The puzzle of dependent measures in experimental pragmatics. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of CogSci 36.*

Degen, J., & Tanenhaus, M. (2019). Constraint-based pragmatic processing. In C. Cummins & N. Katsos (Eds.), *The Oxford Handbook of Experimental Semantics and Pragmatics.* Oxford University Press.

Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Sensorimotor Foundations of Higher Cognition* (pp. 527–574). Oxford University Press.

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers and numerosity perception. *Cognition*, *143*, 115–128.

Donkin, C., Heathcote, A., Brown, S., & Andrews, S. (2009). Non-decision time effects in the lexical decision task. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of CogSci 31*.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314.

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

Grice, H. P. (1975). Logic and conversation. In *Syntax and Semantics: Vol. 3: Speech Acts* (pp. 41–58). New York: Academic Press.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behaviour*, *10*(3), 244–253.

Kaup, B., Zwaan, R., & Lüdtke, J. (2007). The experiential view of language comprehension: How is negated text information represented? In F. Schmalhofer & C. Perfetti (Eds.), *Higher Level Language Processes in the Brain: Inference and Comprehension Processes* (pp. 255–288). Mahwah, NJ: Erlbaum.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633.

Kochari, A. (2019). Conducting web-based experiments for numerical cognition research. *Journal of Cognition*, *2*(1), 39.

Mulder, M., van Maanen, L., & Forstmann, B. (2014). Perceptual decision neurosciences – a model-based review. *Neuroscience*, *277*, 872–84.

Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 316–334.

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218.

Nordmeyer, A. E., & Frank, M. C. (2014). A pragmatic account of the processing of negative sentences. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of CogSci 36.*

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*(3), 446–461.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(1), 873–922.

Ratcliff, R., & McKoon, G. (2018). Modeling numerosity representation with an integrated diffusion model. *Psychological Review*, *125*(2), 183–217.

Schlotterbeck, F. (2017). *From Truth Conditions to Processes: How to Model the Processing Difficulty of Quantified Sentences Based on Semantic Theory*. PhD dissertation, University of Tübingen.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Szymanik, J. (2016). *Quantifiers and Cognition. Logical and Computational Perspectives*. Springer.

Szymanik, J., & Zajenkowski, M. (2013). Monotonicity has only a relative effect on the complexity of quantifier verification. In M. Aloni, M. Franke, & F. Roelofsen (Eds.), *Proceedings of the 19th Amsterdam Colloquium.*

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin Review*, *11*, 192–196.

Waldon, B., & Degen, J. (2020). Modeling behavior in truth value judgment task experiments. In *Proceedings of the Society for Computation in Linguistics* (Vol. 3). (Article 3)