Computational model reveals two sources of polarity across negative adjectival expressions and quantifiers

Fabian Schlotterbeck *[1], Sonia Ramotowska[2], Leendert Van Maanen[3], and Jakub Szymanik[4]

[1]Institute of German Language and Literatures, Universität Tübingen
[2]Institute for Logic, Language and Computation, University of Amsterdam;
[3]Department of Experimental Psychology & Helmholtz Institute, Utrecht University
[4]Center for Mind/Brain Sciences and Dept. of Information Engineering and Computer Science, University of Trento

September 2, 2024

*Corresponding author: fabian.schlotterbeck@uni-tuebingen.de

Contents

1 Introduction 3
1.1 The polarity effect across linguistic expressions 5
1.2 A computational modeling approach 7
1.2.1 Enhancing Interface Transparency with a computational model 8
1.2.2 Chronometric functions and removable interactions 9
1.3 scalinDDM: Diffusion Decision Model application to verification of scalar expressions 10
1.4 Current study 14

2 Methods 15
2.1 Design, materials & procedures 15
2.2 Exclusion criteria 16
2.3 Participants 16
2.4 Data Analysis 16
2.4.1 Regression analysis 17
2.4.2 Model fitting and comparisons 17

3 Results 20
3.1 Descriptive results and regression analysis 21
3.1.1 Interim discussion 24
3.2 scalinDDM analysis: Model fit, posterior distributions and hypothesis tests 24
3.2.1 Exp. 1: proportional quantifiers. 24
3.2.2 Exp. 2: adjectival expressions. 25
3.2.3 Comparison between experiments. 26

4 Discussion 28
4.1 Theoretical accounts of the polarity effect 31
4.2 Interpretation of parameters 32

5 Conclusions 37

References 38

A Prior and posterior distributions 44

Abstract: The polarity effect is one of the most robust effects in linguistics: Negative expressions are processed slower than corresponding positive ones. Decades of research has shown that this effect replicates for different types of negative expressions, however, its magnitude varies depending on the expression type. It has been argued that in addition to polarity, the downward monotonicity increases a cost of processing negative expressions. In this paper, we use computational modeling of data from a sentence-picture verification task to investigate the sources of the polarity effect by contrasting quantifiers (*more than half* vs. *fewer than half*) that differ in polarity and monotonicity with adjectives (*a large proportion* vs. *a small proportion*) that only differ in polarity. We collected reaction times and response data in two web-based sentence-picture verification task experiments. Our reaction time analysis showed a larger polarity effect for quantifiers than adjectives, thus replicating a previously observed non cross-over interaction between polarity and expression type. In order to interpret this interaction with respect to semantic properties of these expressions, we fit scalinDDM, a modified version of the Diffusion Decision Model to the data. We showed that the drift rate and non-decision time parameters were affected by polarity, however, the effect in drift rate was larger for quantifiers than adjectives. This finding suggests there is no exclusive mapping between semantic properties and the model's parameters. Our results show that the computational model can increase transparency in mapping between linguistic properties and dependent variables in an experiment.

Keywords: Diffusion Decision Model; Polarity; Verification task; Quantifiers; Interface Transparency; Computational modeling

1 Introduction

Studies on the processing of negation in the 60s and 70s of last century (e.g. H. H. Clark & Chase, 1972; Just & Carpenter, 1971; Wason, 1961; cf. H. H. Clark, 1976) have started a long-lasting debate on the causes of processing difficulties with negative expressions, as compared to positive ones. The key finding, replicated in many experiments (see Kaup & Dudschig, 2020, for review) [1], is that it takes longer to process a sentence containing a negative expression than a positive one and that the processing of negative expressions is more error-prone. We will refer to this effect as the **polarity effect**. At first glance, this effect may seem not particularly puzzling for sentence negation (also called *explicit* or *overt* negation) because, in languages such as English, this type of negation adds an extra word (*not* or *no*) to the sentence. Therefore, a sentence with negation tends to be longer than one without and is thus also expected to be processed longer given common observations about language comprehension. However, another key finding on negative expressions shows that the polarity effect replicates even when the negation is not explicit (Deschamps et al., 2015; Schlotterbeck et al., 2020; Urbach, DeLong, &

[1] See also Agmon, Loewenstein, & Grodzinsky, 2019, 2022; H. H. Clark & Chase, 1972; Deschamps, Agmon, Loewenstein, & Grodzinsky, 2015; Fischler, Bloom, Childers, Roucos, & Perry, 1983; Kaup, Lüdtke, & Zwaan, 2006; Nieuwland & Kuperberg, 2008; Schlotterbeck, Ramotowska, Van Maanen, & Szymanik, 2020 a.m.o.

Kutas, 2015; Urbach & Kutas, 2010, see Grodzinsky et al., 2020 for discussion). Although the sentence in (1-a) has the same truth conditions and the same number of words as the sentence in (1-b), comparison between sentences like these revealed the same kind of polarity effect as was found with sentence negation (for effects of sentence negation see e.g. H. H. Clark & Chase, 1972; Fischler et al., 1983; Kaup et al., 2006, among many others; see Kaup & Dudschig, 2020, for review).

(1) a. More than half of the students passed the exam.
b. Fewer than half of the students failed the exam.

With regard to their meaning, polar opposite expressions like *more than half* vs. *fewer than half* are mirror images that can be related to each other via an operation of scale reversal. [2] Negative polarity is then usually defined via markedness. Positive polar expressions are unmarked and preferred whereas negative expressions are marked and dispreferred (Just & Carpenter, 1971, cf. Greenberg, 1963). This can be seen clearly in questions like the following.

(2) a. How many students passed the exam?
b. ?How few students passed the exam?

The question in (2-a) is the intuitively preferred way to ask about the number of students who passed, whereas (2-b) is marked and dispreferred. Related to this, positive expressions also have a neutral meaning in questions whereas negative expressions do not. The question in (2-a) is a neutral question about the number of student, while the question (2-b) allows a pragmatic inference that the number was rather small (see van Tiel & Pankratz, 2021 for a systematic approach to polarity classification integrating various criteria).

The polarity effect has been tested by using time-sensitive methods to monitor how the interpretation of negative expressions changes over time. Various measures indicate that the initial interpretation of a negative expression may differ from the interpretation that is arrived at later on. These measures include choice reaction times (RT) measured after various onsets of the presentation of a stimulus (e.g. a picture that has to be verified after having read a negated sentence; Kaup et al., 2006), ERPs measured during reading and contrasted with offline judgments (Nieuwland, 2016; Urbach et al., 2015; Urbach & Kutas, 2010), eye-movements across a trial in visual-world eye-tracking experiments (Tian, Ferguson, & Breheny, 2016; Vanek, Matic Škoric, Košutar, Matějka, & Stone, 2024) or mouse-tracking in combination with MEG analyses (Zuanazzi et al., 2024).

The polarity effect has been also investigated by varying the context in which a negative expression is processed. This could be the local context within a sentence (e.g. clefted vs. un-clefted versions of a sentence, Tian, Breheny, & Ferguson, 2010, or a subordinate clause Nieuwland & Kuperberg, 2008), the discourse context (Urbach et al., 2015) or some visual context (Nordmeyer & Frank, 2014; Xiang, Kramer, & Nord-

[2] A scale is an ordered set (Solt, 2015a). By scale reversal, we refer to a mapping that reverses this order (e.g. the mapping from the natural numbers to their negative inverses, $n \mapsto -n$, extending their natural ordering, $\leq$).

meyer, 2020). These studies found that processing difficulty of negative as compared to positive expressions can be reduced or even completely alleviated in supportive contexts or in contexts that exhibit a high degree of predictability (but see also Augurzky, Schlotterbeck, & Ulrich, 2020; Rück et al., 2021; Xiang et al., 2020; for studies that found the polarity effect even with contextual embedding exhibiting such a high degree of predictability).

A number of explanations for the polarity effect have been given in the literature. Some of them are derived from psycholinguistic theoriess (e.g., the two-step model; see H. H. Clark, 1976; H. H. Clark & Chase, 1972; Kaup et al., 2006); others are based on the semantic and pragmatic properties of the expressions under investigation, without explicit reference to a processing model (e.g. Agmon et al., 2019) and some use a combination of these two approaches (e.g. Bott, Schlotterbeck, & Klein, 2019; Nordmeyer & Frank, 2014; Tian et al., 2010; Xiang et al., 2020). However, the results of experimental studies are inconclusive concerning which approach is superior. For example, it is unclear if the delay in processing negative expressions is due to higher complexity of the semantic representation of negative content or due to more complex verification procedures (see e.g. Bott et al., 2019; Grodzinsky, Agmon, Snir, Deschamps, & Loewenstein, 2018, for discussion). One particular problem in this regard is that it is difficult to isolate the relevant linguistic properties of the expressions unambiguously. The polarity of an expression correlates with other semantic and pragmatic properties, for example monotonicity or focus patterns, which may mediate the magnitude of the polarity effect (Agmon et al., 2019; Just & Carpenter, 1971).

1.1 The polarity effect across linguistic expressions

While the polarity effect has been observed in various languages (e.g. Agmon et al., 2022; Kaup et al., 2006) and across different types of negative expressions, negative expressions are not a uniform class. Some studies showed that the magnitude of the polarity effect measured by reaction times differs across linguistic expressions. One straightforward way to test how semantic and pragmatic properties of linguistic expressions contribute to the polarity effect, is to find suitable contrasts between expressions that tease apart the properties of interest. For example, using a verification task, Just and Carpenter (1971) compared the polar opposite quantity words *many* vs. *few* (as in (3); cf. also the examples in (1) containing the same quantity words in comparative form) to cases like (4). They called expressions like in (3-b) and (4-b) *syntactic* and *semantic negatives*, respectively.

(3) a. Many of the students passed the exam.
    b. Few of the students failed the exam.

(4) a. A large proportion of the students passed the exam.
    b. A small proportion of the students failed the exam.

Both types of negatives are marked and dispreferred. Just and Carpenter (1971) found that both are also processed slower. They argued furthermore that syntactic negatives have additional sources of processing difficulties due to a shift in focus. Semantic negatives focus on smaller quantities. For example, *a small proportion* in (4) focuses on the quantity of the students who failed the exam and asserts that this quantity is small in proportion to all students in the comparison class. In contrast, the syntactic negatives focus on the larger quantity (see also Moxey & Sanford, 1986; Moxey, Sanford, & Dawydiak, 2001 for related focusing effects in discourse processing). According to Just and Carpenter's (1971) proposal, the expression *fewer than half* in (1) is represented as a negation of its positive counterpart *more than half* (roughly: *It is not the case that more than half ...*). Therefore, the syntactic negatives like positives focus on a larger quantity but in contrast to positives they deny a certain property about this quantity. In other words, *more than half* asserts that the larger proportion of students passed the exam and *fewer than half* denies that the larger proportion of students failed the exam. In support of their proposal, Just and Carpenter (1971) showed that varying the ordering between sentence and picture presentation in verification tasks affects the two types of negatives differently. If the sentence is presented first, it guides focus, but if the picture is presented first there may be a mismatch between picture- and sentence induced focus.

In addition to polarity, some negative expressions differ with respect to their monotonicity properties. Monotonicity refers to entailment patterns licensed by linguistic expressions. Upward monotone expressions license inferences from subsets to supersets. For example, the sentence "My students live in Amsterdam." entails that "My students live in the Netherlands." Upward-entailing quantifiers such as *more than half* license the same type of inference: "*More than half* of my students live in Amsterdam." entails that "*More than half* of my students live in the Netherlands." Negation, however, reverses the inference pattern. The sentence "My students do not live in the Netherlands." entails that they do not live in Amsterdam. The negative quantifier *fewer than half* licenses the same inference as sentence negation: "*Fewer than half* of my students live in the Netherlands." entails "*Fewer than half* of my students live in Amsterdam." Therefore, we say that a quantifier like *fewer than half* is downward monotone.

In general, positive expressions tend to be upward monotone and negative expressions tend to be downward monotone. However, this association is not perfect as there are also negative expressions that are not downward monotone. This dissociation between negativity and monotonicity is in fact displayed by negative expressions like *a small proportion* or *a small number.* That these expressions are not downward monotone can be seen in the following contrasts.

(5) a. To pass the exam, you should make few mistakes.
b. #To pass the exam, you should make a small number of mistakes.
(from Agmon et al., 2019)

(6) a. To pass the exam, you should get few answers wrong.
b. #To pass the exam, you should get a small proportion of answers wrong.

We assume with Agmon et al. (2019) that (5-b) and (6-b) sound odd because the empty set is not in the denotation of *a small number of mistakes* or *a small proportion of* *answers*. Since the empty set is a subset of every set, it follows that *a small number* is not downward monotone, i.e. inference from superset to subset is not licensed in general. Agmon et al. (2019) argued that it is this downward monotonicity that adds additional processing cost to expressions like *fewer than half*, which are both negative and downward monotone.

Agmon et al. (2019) contrasted quantifiers like (1) that differ in polarity and monotonicity with adjectives like in (4) that differ only in polarity. They showed that the magnitude of the polarity effect differs between case like in (1) and (4). They showed that the reaction times difference between pairs of "quantifiers", as in (1), was larger than the difference between pairs of "adjectives", as in (4). In addition, the difference between negative quantifiers and adjectives was also found in stronger brain activation during processing of the former ones than the latter Agmon, Bain, and Deschamps (2021). Agmon et al. (2019) therefore suggested that some negative expressions are more negative than others.

To summarize, the studies of Just and Carpenter (1971) and Agmon et al. (2019) indicate that polarity is not a simple dichotomy but rather a gradient or multidimensional concept (see also Brasoveanu, Clercq, Farkas, & Roelofsen, 2014; Denić, Homer, Rothschild, & Chemla, 2021). While negative expressions in general seem to be more difficult to process than positive ones, the magnitude of this cost varies between different types of expressions. Additional processing costs have been attributed to properties that correlate with negative polarity, such as downward monotonicity or focus patterns.

1.2 A computational modeling approach

One general limitation common to all the mentioned studies is that up until today we are lacking an explicit processing model linking theoretical assumptions directly with the dependent variables in the experiments. Therefore, it is often difficult to relate effects in the dependent measures back to theoretical concepts. Providing such a link is exactly what we aim to do in our current approach.

In the current study, we combined the approach of Agmon et al. (2019) with explicit computational modeling of choice reaction times and proportions of errors in order to see how exactly the polarity effects come about in different types of negative expressions. We investigated the sources of the polarity effect in two different types of expressions: quantifiers (*more than half* vs. *fewer than half*) and adjectives (*a large proportion* vs. *a small proportion*), as in (1) and (4). [3] To identify the source of processing difficulties across these expressions, we applied a generative computational model of decision making to precisely quantify the contribution of different sources of the effects in verification reaction times and errors.

[3] We adopt the terminology from Agmon et al., 2019 for simplicity here but note that it does not refer to the properties that actually distinguish between the tested types of expressions. In fact, both contain adjectives (*many* vs. *few* and *large* vs. *small*) and are used as quantifiers.

1.2.1 Enhancing Interface Transparency with a computational model

On a general note, we argue that using an explicit model of choice RT and proportions of errors provides us with a transparent interface between linguistic representations and the decision processes that are based on these representations, even beyond what previous approaches were able to establish based on proportions of errors alone (e.g. Knowlton, Pietroski, Halberda, & Lidz, 2022; Lidz, Pietroski, Halberda, & Hunter, 2011; Pietroski, Lidz, Hunter, & Halberda, 2009; cf. these studies also with Bader & Haeussler, 2010; Huang & Ferreira, 2020, for a similar approach in the domain of syntactic acceptability and gramaticality judgments). Experimental data can inform linguistic theories not only because they are more reliable than pure introspection, the classical method used in traditional linguistics, but also because they can provide additional information about the cognitive processes underlying these representations and their processing trajectory in the form of e.g., reaction times, event related potentials, or eye-movements. However, interpreting this information requires linkage assumptions between the dependent measure in an experiment and abstract concepts from linguistic theory, which can usually not be measured directly.

In the field of experimental semantics, Lidz et al. (2011, see also Pietroski et al., 2009) proposed a convincing linkage assumption under the label of Interface Transparency. They studied the meaning of the superlative proportional quantifier *most* and noted that evaluating the truth value of a simple sentence like *most of the dots are blue* is closely related to numerosity comparison, i.e. the task of comparing the cardinalities of two sets, which has been studied intensively in psychophysics (Dehaene, 2007; Dehaene, Bossini, & Giraux, 1993; Feigenson, Dehaene, & Spelke, 2004; Kaufman, Lord, Reese, & Volkmann, 1949). Lidz et al. (2011) studied verification of the quantifier *most* using methods from psychophysics to derive predictions of error rates given different truth-conditionally equivalent semantic representations of *most.* To this end, they used so-called *psychometric functions*, which predict error rates, a measure of task difficulty, from the numerosities involved in the verification task. In this way, they were able to explicitly link the dependent variable (error rates) with competing but truth-conditionally equivalent semantic representations of *most*. This allowed them to identify one out of several candidate representations based on their experimental results.

Their work on *most* led Lidz et al. (2011) to propose the Interface Transparency Thesis (ITT) which posits a transparent interface between semantic representations and verification procedures as implemented in potentially different cognitive modules (e.g. involving the number system). As they put it, verification procedures are biased towards the operations that are directly encoded in the semantic representations. They provided evidence for the ITT by demonstrating that the procedure used to verify *most* stayed constant across various experimental conditions even if these favored a different representation.

Our current approach is inspired by the ITT but goes beyond it because we are also able to model RT in addition to error rates. In particular, we specify *chronometric functions*, mapping latent theoretical variables to RT, in addition to psychometric functions as used by Lidz et al. (2011) to map properties of their stimuli to proportions of errors.

We achieve this by using the Diffusion Decision Model (DDM; a so-called a sequential sampling decision model that is described in section 1.3 below), which is a generalization of Signal Detection Theory that takes into account the time spent on a decision in addition to the final outcome (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006). The DDM models the joint distribution of RTs and responses for each choice option in a two alternatives force choice (2AFC) tasks. It has been applied successfully to a large variety of such tasks and it explains many common findings e.g. speed-accuracy trade-offs or skewed RT distributions (for review, see e.g. Mulder, van Maanen, & Forstmann, 2014; Ratcliff, Smith, Brown, & McKoon, 2016). The model has a number of free parameters, i.e. latent variables, with a transparant cognitive interpretation. In our current application of the DDM, we build on previous applications of the model to quantifier verification which have already established a transparent interpretation for some key model parameters in terms of notions from linguistic theory (Ramotowska, Steinert-Threlkeld, van Maanen, & Szymanik, 2023; Schlotterbeck et al., 2020). As predicted by the ITT, these previous studies found evidence for verification procedures that are stable across different tasks (Potthoff, Ramotowska, Szymanik, & van Maanen, 2023) and also across time.

Here we argue that modeling errors and RTs jointly gives us an **enhanced interface transparency** in the sense that we can disentangle different components of the verification procedure that affect RTs and errors together. Thus, our approach improves on the ITT because the interface we uncover is transparent in both directions: As predicted by the ITT, we find evidence that linguistic representations bias verification processes, but in addition the current approach unveils also a transparent interface between the composition of these verification processes and the dependent measures in our experiments. Before we describe our computational model below, we relate back to the polarity effect across different types of expressions and discuss the specific motivation of the current study.

1.2.2 Chronometric functions and removable interactions

While Agmon et al. (2019) observed larger effects of polarity in RT for quantifiers than for adjectives, negative polarity still led to delays in both types of expressions. Thus, the interaction they found is not a cross-over interaction. Such non-cross-over interactions have been called "removable" because they can be removed by monotonic transformations of RT (Loftus, 1978; Wagenmakers, Krypotos, Criss, & Iverson, 2012). This is of central importance because, as is typical, Agmon et al. (2019) were not interested in RT *per se* but rather in RT as a means to infer a latent, i.e. unobservable, variable characterizing the underlying cognitive process. In particular, their main hypothesis concerned the effects of polarity (positive vs. negative) and expression type (quantifiers vs. adjectives) on "cognitive cost". However, without knowing how exactly RT maps to the underlying latent variables such as the cognitive difficulty of the task (for definitions of this and other relevant latent variables see section 1.4, where we introduce our computational model), removable interactions like the one they found do not allow reliable conclusions about these latent variables. To illustrate our main motivation, we use the

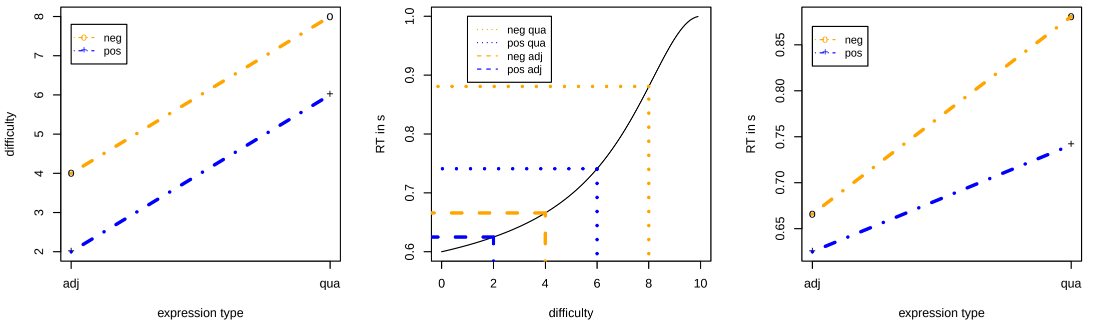


Figure 1: Left: Interaction plot showing additive effects of expression type and polarity on task difficulty. Middle: Hypothetical mapping from task difficulty (in an arbitrary unit) to RT based on the chronometric function derived by Palmer et al. (2005) from a type of sequential sampling decision model similar to the one used in the current study. Right: Hypothetical interaction plot for the scenario on the left and compatible with the findings from Agmon et al. (2019). The abbreviations *pos*, *neg*, *adj* and *qua* stand for positive, negative, adjective and quantifier, respectively.

non-linear but nevertheless strictly monotone chronometric function shown in Figure 1, mapping from task difficulty to RT. This specific function has been shown to fit mean RT well in perceptual decision tasks (Palmer, Huk, & Shadlen, 2005) and is, moreover, derived from the same type of decision model that we apply also in the current approach. For the case at hand, i.e. the verification of proportional quantifiers as in (1) and (4), task difficulty is affected by the proportion of objects with the relevant property (e.g. the proportion students that passed the exam), polarity (Deschamps et al., 2015), and expression type (Agmon et al., 2019). Using the chronometric function from Palmer et al. (2005), Figure 1 shows that the type of interaction Agmon et al. (2019) observed in mean RT could have emerged from purely additive effects of expression types and polarity on task difficulty. In the type of scenario shown in Figure 1, the extra "cognitive cost" Agmon et al. (2019) measured for the downward entailing quantifiers would be solely due to the shape of the chronometric function, and not to effects in task difficulty itself. In order to distinguish this scenario from other potential sources of the observed effects, we used a well-established explicit model of choice RT akin to the model underlying the chronometric function in Figure 1 and tested directly for interactions in the underlying variables themselves. This allows us to draw more reliable conclusions about the underlying cognitive processes.

1.3 scalinDDM: Diffusion Decision Model application to verification of scalar expressions

In the DDM, the decision options are conceptualized as two decision boundaries and the process of decision-making is a noisy process of evidence accumulation toward one

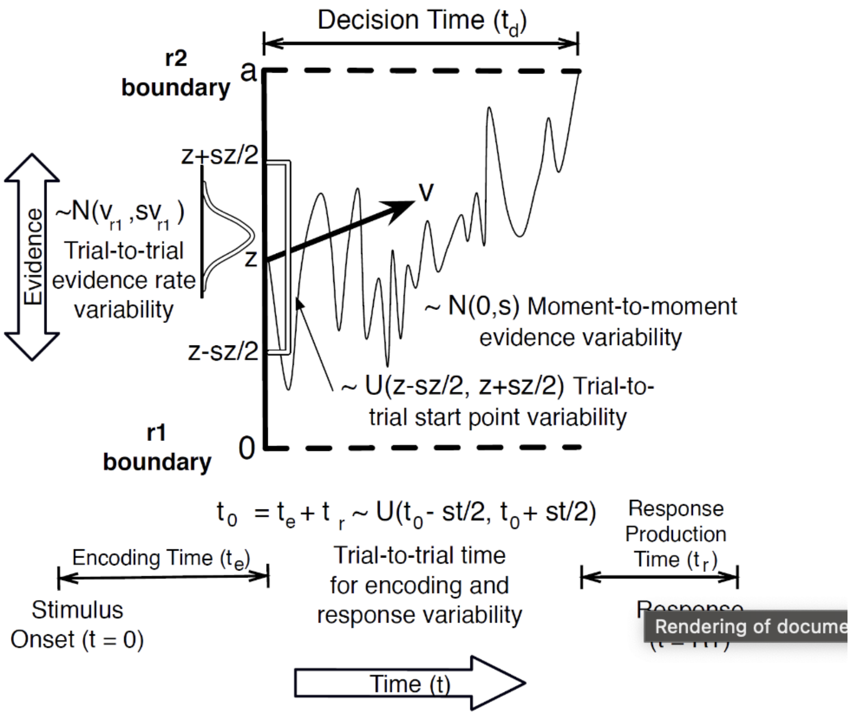

Decision Time (t_d)
r2 boundary
a
z+sz/2
Evidence
~N(v_{r1} ,sv_{r1})
Trial-to-trial evidence rate variability
z
v
~ N(0,s) Moment-to-moment evidence variability
z-sz/2
~ U(z-sz/2, z+sz/2) Trial-to-trial start point variability
r1 boundary
0
t_0 = t_e+ t_r ~ U(t_0- st/2, t_0+ st/2)
Trial-to-trial time for encoding and response variability
Response Production Time (t_r)
Encoding Time (t_e)
Stimulus Onset (t = 0)
Response
Rendering of docume
Time (t)

Figure 2: Graphical overview of the DDM taken from Heathcote et al. (2019). The two response alternatives are denoted by r1 and r2 and separation between the corresponding boundaries by a. Parameters are explained in the text in the current section and in section 2.4.2.

of these boundaries. Thus, it models decision making as a stochastic process. The process of accumulating evidence could for example refer to the integration of perceptual information (e.g., about the number of stimuli on a display) or retrieval of internal representation from memory (e.g., in the lexical decision task). The model has a number of free parameters that account for the speed of the decision process (drift rate), time for stimuli encoding and response execution (non-decision time), initial decision bias (starting point), and amount of evidence needed before a decision is made (boundaries separation). In particular, the drift rate parameter captures the evidence accumulation process. Its value is thus dependent on the quality of the accumulated information (e.g., quality of the perceptual stimuli) or the quality of the internal representation (e.g., accessibility in memory). Figure 2 provides a graphical overview of the model components and corresponding parameters.

Verification of quantity expressions can be conceptualized as a 2AFC task and can be modeled with the DDM. The two decision boundaries correspond to the truth-value judgment options (true or false) and the evidence accumulation process is a process of comparison of the meaning representation of a quantity expression to the context against which it is evaluated (the *verification scenario*). The verification scenario contains numerical information either given as a precise number (e.g., percentage) or visually depicting different numerosities of objects (e.g., 10 blue and 11 yellow dots). In this way, verification of quantity expressions is an analog of the number comparison task, which has been already modeled with the DDM (Kang & Ratcliff, 2020; Ratcliff & McKoon, 2018, 2020). For example, Ratcliff and McKoon (2018) integrated the approximate number representations into the DDM's drift rate because this parameter represents the quality of the stimuli representation.

Ramotowska et al. (2023) and Schlotterbeck et al. (2020) adapted the DDM to model semantic representations of quantity expressions. We call this model the scalinDDM (for *scalar language induced diffusion decision model*). According to this model, participants' truth-value judgments in the verification task depend on the semantic representation of a quantifier and numerical information from a verification scenario. scalinDDM implements the semantic representations of quantifiers in the drift rate parameter (cf. approach of Ratcliff & McKoon, 2018). Concretely, scalinDDM uses the generalized logistic regression function as a link function between participants' representations of quantifier meanings (see Figure 3) and verification scenarios on the one hand and drift rates in the verification task on the other. This link function is governed by four parameters, namely a midpoint parameter that determines where negative drift rates switch to positive, a steepness parameter that determines how fast drift rates increase or decrease with distance from the midpoint, and upper and lower asymptote parameters which determine the maximum possible drift rates towards the two decision boundaries.

The first two parameters, midpoint and steepness, specify the truth-conditional representations of quantifiers. The truth value of a quantified sentence is determined by a threshold, which corresponds to the midpoint of the logistic function. The upward entailing quantifiers are true above the threshold and the downward entailing quantifiers are true below the threshold [4]. In other words, the threshold determines the required quantity at which the truth value of the quantifier switches. For example, for quantifiers such as *more than half* or *fewer than half* this threshold is a 50% proportion (cf. (7)), while other quantifiers *most*, *few* and *many* can have variable and context dependent thresholds.

(7) Truth-conditional representations of quantifiers:

a. *More than half*$(A, B) = 1$ iff $\mid A \cap B \mid > \frac{|A|}{2}$

b. *Fewer than half*$(A, B) = 1$ iff $\mid A \cap B \mid < \frac{|A|}{2}$

Moreover, scalinDDM allows us to capture uncertainty about the threshold parameter with the steepness parameter. The uncertainty may come from the quantifier meaning. For example, some quantifiers, such as *many* or *few* are vague (see e.g. Solt, 2015b) and for these quantifiers participants make slower decisions for proportions close to the

[4] See van Tiel, Franke, and Sauerland (2021) for a similar implementation in a pragmatic computational model of language production.

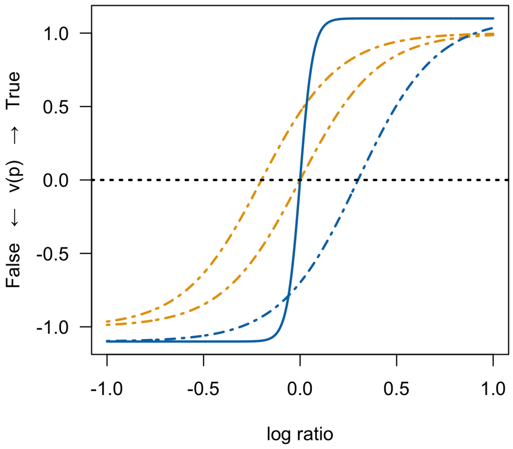


Figure 3: Four types of drift rate (v(p)) indicated by four lines. The log ratio for which v(p) = 0 is the midpoint. The line types indicate different values of the vagueness parameter: dashed lines indicate drift rate of vaguer quantity expressions (lower values of drift rate around the threshold and less steep lines). The colors indicates different values of the asymptote parameters: blue indicates more extreme values of asymptotes (higher drift rate).

threshold. The uncertainty may also interact with the numerical information that can be extracted from the verification scenario. For example, if the numerical information in the verification scenario is given as a precise proportion (e.g., 67%) the uncertainty is smaller than if the information is imprecise (e.g., the ratio between two types of objects on the visual display).

In addition to modeling the differences in vagueness and truth conditions of linguistic expressions, scalinDDM allows also for testing the effect of polarity on model parameters. Schlotterbeck et al. (2020) used scalinDDM to study the polarity effect in the polar opposite quantifiers *more than half* and *fewer than half.* They found that two parameter values differed between them. The non-decision time was longer for the negative quantifier than the positive one. Moreover, they used the distance between the two asymptotes as a measure of maximum speed of verification and they showed that this distance is greater for the positive quantifier than the negative one. The study, therefore, did not reveal a unique and unequivocal answer to the question about the source of the polarity effect. We intend to investigate this question further in the current paper by applying scalinDDM to the comparison between polar opposite quantifiers (*more than half* vs. *fewer than half*) and adjectives (*a large* vs. *a small* proportion) that we adopted from Agmon et al. (2019).

To this end, we make use of a straightforward extension of scalinDDM to the verification of adjectives. The truth-conditional representations for adjectives such as *a large proportion* and *a small proportion* can be formulated as in (8), where $k_{large/small}$ are fractions.

(8) Truth-conditional representations of adjectives:
   a. $A\ large\ proportion(A, B) = 1$ iff $\mid A \cap B \mid > \mid A \mid \times k_{large}$
   b. $A\ small\ proportion(A, B) = 1$ iff $\mid A \cap B \mid < \mid A \mid \times k_{small}$

The midpoint parameter of the drift rate function can capture the difference in semantic representations between quantifiers and adjectives in terms of variability of the thresholds. The threshold for quantifiers is predicted to be fixed at 50% whereas the threshold for adjectives can vary between participants and also between positive and negative adjectives. The steepness parameter can further capture the difference in vagueness between expression types. In this way, scalinDDM makes it possible to separate different sources of variability in responses and reaction times that come from semantic differences between expressions not related to the polarity effect.

1.4 Current study

The goal of the current study was to disentangle different sources of the polarity effect. We intend to investigate these sources by applying scalinDDM to the comparison between polar opposite quantifiers (*more than half* vs. *fewer than half*) and adjectives (*a large* vs. *a small* proportion) that we adopted from Agmon et al. (2019). We analysed responses and reaction times data from two sentence-picture verification experiments. In Experiment 1, participants read sentences involving polar opposite quantifiers as in

(9-a/b) and in Experiment 2 they read sentences of the same structure but with the adjectives in (9-c/d). Participants verified these sentences based on pictures presenting two sets of dots in two different colors.

(9) a. *More than half* of the dots are blue. (MTH)
b. *Fewer than half* of the dots are blue. (FTH)
c. *A large proportion* of the dots are blue. (LPROP)
d. *A small proportion* of the dots are blue. (SPROP)

We expected to replicate the main results of Agmon et al. (2019), especially the non-crossover interaction discussed above. In terms of scalinDDM parameters, we expected to replicate the findings of Schlotterbeck et al. (2020) that drift rates and non-decision times may be affected by polarity. Regarding the source of the expected interaction in RT, we distinguish several possible scenarios. The first possibility is that both downward monotonicity and negative polarity are separate sources of difficulty that map onto distinct parameters of scalinDDM. Under this scenario, we expect effects in both non-decision time and drift rates for the quantifiers (MTH VS. FTH) but only an effect in one of these parameter in the adjectives (LPROP VS. SPROP) because they differ only in polarity and not in monotonicity. A completely different scenario would be one where there are only purely additive effects in the parameters of scalinDDM and the interaction is simply due to a non-linear mapping between drift rates and RT (cf. Figure 1 and discussion above). This is the scenario in which there would be no real interaction in task difficulty. Yet another scenario would be one where we observe an interaction in one of the two parameters but only additive effects in the other one. Such a result could, for example, be due to monotonicity and polarity jointly affecting one of the two parameters but only one of these properties affecting the other parameter.

2 Methods

2.1 Design, materials & procedures

Participants first read a complete sentence like in (9) above, 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 POLARITY (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 condition, 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 minutes on average.

Exp. 2 was identical to Exp. 1 except for the fact that the adjectival expressions LPROP vs. SPROP were tested instead of MTH vs. FTH (cf. example in (9)). The procedure, instructions, number of conditions and trials were all identical. We thus manipulated the factor EXPRESSION TYPE (levels: *quantifiers* vs. *adjectives*) as a third factor between the participants of Exp. 1 and Exp. 2.

2.2 Exclusion criteria

In Exp. 1, the following criteria were used to exclude 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. All of the latter had, however, already been excluded by one of the other criteria. We excluded trials with reading times or RT shorter than 200 ms or longer than mean+3.5*SD (calculated per participant and condition). In Exp. 2, the accuracy-based criterion was replaced by a criterion related to the sensitivity of participants' judgments to log ratios.

2.3 Participants

Participants were recruited via `prolific.co`. For Exp. 1, testing the quantifiers MTH vs. FTH, 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 years; $sd = 13$; range: $18 - 69$, compensation £7.5). For Exp. 2, testing the quantifiers LPROP vs. SPROP, data from 98 English native speakers was collected in total and after exclusion the final sample consisted of 64 participants (39 female; mean age 38 years; $sd = 13$; range: $16 - 65$, compensation £7.5).

2.4 Data Analysis

Exps. 1 & 2 were first analyzed using mixed-effects regression models. The motivation for these analyses was to test whether previous findings were replicated by our results. Afterwards, we applied scalinDDM to the data from both experiments in order to test our hypotheses above the cognitive sources of the observed effects.

2.4.1 Regression analysis

We report a combined mixed-effects regression analysis of log-transformed RTs from Exps. 1 & 2 and a separate analysis of errors from Exp. 1. [5] Most importantly, we tested for the interaction that motivated our study, i.e. the interaction between EXPRESSION TYPE and POLARITY in RT. In addition, we tested for three other known types of effects: distance effects, where smaller absolute ratios lead to increased difficulty, POLARITY effects, and interactions between POLARITY and truth value (e.g. Agmon et al., 2019; Just & Carpenter, 1971), where the effect of POLARITY is larger in true than in false conditions. For the purpose of these analyses, independent variables were recoded in the following way. The absolute value of the logarithm of the ratio of the two presented numerosities in each trial (ABSOLUTE LOG RATIO) was included as a factor (levels: .167 vs. .336). In addition, the factors POLARITY (levels: *positive* vs. *negative*) and TRUTH VALUE (levels: *true* vs. *false*) were included in the analysis of errors from Exp. 1. Conditions with *mth* were coded as *true* if log ratio was positive and as *false* if they were negative. For *fth*, TRUTH VALUE was coded the opposite way. For Exp. 2, accuracy was not part of the regression analysis because vague adjectives like *large* and *small* do not allow for unequivocal truth-value judgments. For the same reason, we also did not include the factor TRUTH VALUE in the combined regression analysis of RT but instead included the degree of APPROPRIATENESS (levels: *high* and *low*) which was however encoded in exactly the same way as TRUTH VALUE. The combined analysis of RT also included the factors POLARITY and ABSOLUTE LOG RATIO. In addition, EXPRESSION TYPE (*adjectives* vs. *quantifiers*) was included as between-participants factor.

2.4.2 Model fitting and comparisons

Strategy: We used a hierarchical Bayesian model fit procedure. One advantage of the Bayesian hierarchical approach in contrast to frequentist approaches was that we could handle quasi complete separation (see e.g. R. G. Clark, Blanchard, Hui, Tian, & Woods, 2023 for related discussion) that emerged because of individual differences with respect to adjective thresholds (i.e. what individual considered large or small in the context of the experiment). We used the following strategy to fit scalinDDM to the data from Experiments 1 & 2: First, a Bayesian hierarchical version of the DDM (implemented in the Dynamic Models of Choice, DMC software Heathcote et al. (2019)) was fit to the data set obtained from the comparative quantifiers (Exp. 1) using Markov Chain Monte Carlo (MCMC) simulation. To this end, we adopted parameter transformations and constraints from Schlotterbeck et al. (2020; see below) that had been determined and evaluated using model comparisons based on (partly overlapping) data from highly comparable verification experiments but using a frequentist, non-hierarchical approach. Prior distributions for the current Bayesian analysis were informed by the systematic

[5] We applied the log-transformation to RT in order to maximize comparability with Agmon et al. (2019), who used the transformation in order to meet the distributional assumption of the of linear mixed models. Note that the log-transformation itself is a non-linear but monotonic transformation that may affect the presence or absence of interactions.

parameter review of Tran, van Maanen, Heathcote, and Matzke (2021, see Figure **??**). If posterior distributions of free model parameters in the current analysis did not differ significantly between positive and negative POLARITY, as judged by their 95% credible intervals (CIs), we added a further constraint, forcing them to be constant across POLARITY. Next, we chose the better of the two models based on a Bayes factor (e.g. Nicenboim, Schad, & Vasishth, 2021) comparing the constrained with the unconstrained model. The winning model was then fitted to the data set obtained from the adjectival expressions (Exp. 2). Two further constraints were implemented based on *post hoc* observations and again tested using model comparisons. These are described in the following subsection.

Parameter constraints and prior distributions The relationship between log ratios, $\log r$, and drift rates, $v(\log r)$, was specified using a generalized logistic function (cf. Ramotowska et al., 2023; Schlotterbeck et al., 2020):

$$v(\log r) = V_l + \frac{V_l - V_u}{1 + \mathrm{e}^{-g(\log r - \log r_0)}}, \tag{1}$$

where V_l and V_u are the lower and upper asymptote, respectively, g is the growth rate and $\log r_0$ is the midpoint, i.e. that log ratio for which drift rate is predicted to be 0. (see Figure 3 above for illustration). Specifying drift rates in this way introduced an additional set of free model parameters, increasing their total number in comparison to the standard DDM by three if we consider one participant in one experimental condition separately. At the same time though, the logistic drift-rate function acted as a theoretically motivated constraint on the model because drift rates were forced to obey this strictly monotonic relationship with log ratios across all experimental conditions (cf. Ratcliff & McKoon, 2018). Thus, for the entire experiments the numbers of parameters were reduced by four per participant (ignoring hierarchical structure for the moment).

In addition, we applied the following parameter transformation to starting points: In the DMC software the starting point of the diffusion process, z, is defined relative to the boundary separation, a, and ranges between 0 and 1. We applied the logit transformation to relative starting points and thus estimated $\mathrm{logit}(z)$ instead of z. The motivation for this transformation was to eliminate the upper and lower bounds on possible starting points in order to make the range of possible starting points compatible with the support of normal prior distributions.

The midpoints were not constrained in the present study because we expected that the vague adjectival meanings investigated in Exp. 2 would require variability between positive and negative expressions. The basis for this expectation was that midpoints correspond to ‘pragmatic thresholds’ for accepting vague expressions like *large* or *small* as true (see section 1.3 above for explanation) and these have been shown in previous research to vary between positive and negative expressions (Ramotowska, Steinert-Threlkeld, Van Maanen, & Szymanik, 2020; Ramotowska et al., 2023; Schoeller & Franke, 2015).

In our initial model, we allowed asymptotes (V_l and V_u) and midpoints ($\log r_0$) of the drift rate function as well as non-decision times (t_0) and logit-transformed starting points to vary with POLARITY, whereas boundary separation (a), growth rate (g) of the drift rate function as well as all variability parameters were constrained to be constant across positive and negative POLARITY. However, as mentioned already in the previous section, two further constraints were added based on *post hoc* observations: Firstly, we noticed after the just described model had been fitted to the verification data obtained from the comparative quantifiers in Exp. 1 (using the prior distributions described in the following paragraph) that the absolute values of the upper and lower asymptotes did not differ from each other because they were symmetric around zero for both quantifier types. Therefore, we fitted a simpler model in which the upper and lower asymptotes were constrained to the negative inverse of each other. We compared these two models by computing their Bayes factor. A Bayes factor of 52720 revealed that the likelihood of the simpler, more constrained model exceeds that of the more complex by a factor of more than 50000. Secondly, we noticed that in the winning models the starting point parameter (log-transfomed z) showed a significant negative correlation with the midpoint parameter r_0, within positive and negative expressions. For larger mid points, indicating a later switch from negative to positive drift rates, there was a stronger bias towards the response that corresponds to the negative drift. We thus observed a trade-off between these two parameters, and they were the only pair of parameters showing such a trade-off. Since posterior distributions for the starting point parameter clustered closely together at values slightly above .5 (95% CIs between .516 and .550), we decided to constrain them to be equal across POLARITY as well. A comparison between the constrained and unconstrained model favored the constrained model by a Bayes factor of 66702. We therefore report this final constrained model below (but see also Appendix **??** for the results based on the model with unconstrained starting points and a description of correlations among all pairs of model parameters).

As is common practice in DDM analyses, the variability parameters s and s_{t0} were set to the constants 1 and 0, respectively, and d, which corresponds to differences in non-decision time between the two response alternatives was set to 0, amounting to no difference at all. In this study, we allowed for trial-to-trial variability in starting points and drift rates, i.e. s_z and s_v were allowed to differ from zero, but were still constrained across POLARITY. An overview of the final model parameters and constraints on them is given in Table 1.

In total, the final model had twelve parameters. Four of them (V_u, s, s_{t_0} and d) were constrained to specific values and, of the remaining eight, four were allowed to vary across POLARITY. Thus, eleven free model parameters were estimated during model fitting, once for each participant and twice on the level of hyper parameters, namely for hyper means and hyper standard deviations. The 33 prior distributions for all these parameters are shown in Figure 11 and specified in Table 2, which also includes upper and lower bounds on possible values. On the participant level, truncated normal distributions were used, as is required by the DMC software. If possible, the location and scale parameters (p_1 and p_2) were taken directly from the distributions derived by Tran et al. (2021), or from its dominant part in case they proposed mixture distributions. In cases where Tran et

symbol	description	constraint
t_0	non-decision time	–
$\log(r_0)$	midpoint of $v(\log r)$	–
V_l	lower asymptote of v	–
V_u	upper asymptote of v	constrained to $-V_l$
$\text{logit}(z)$	transformed starting point	constrained across POLARITY
g	growth rate of $v(\log r)$	constrained across POLARITY
a	boundary separation	constrained across POLARITY
s_z	trial to trial variability in z	constrained across POLARITY
s_v	trial to trial variability in $v(\log r)$	constrained across POLARITY
s_{t_0}	trial to trial variability in t_0	constrained to 0
d	variability in t_0 between response alternatives	constrained to 0
s	moment to moment variability in $v(\log r)$	constrained to 1

Table 1: Parameter constraints

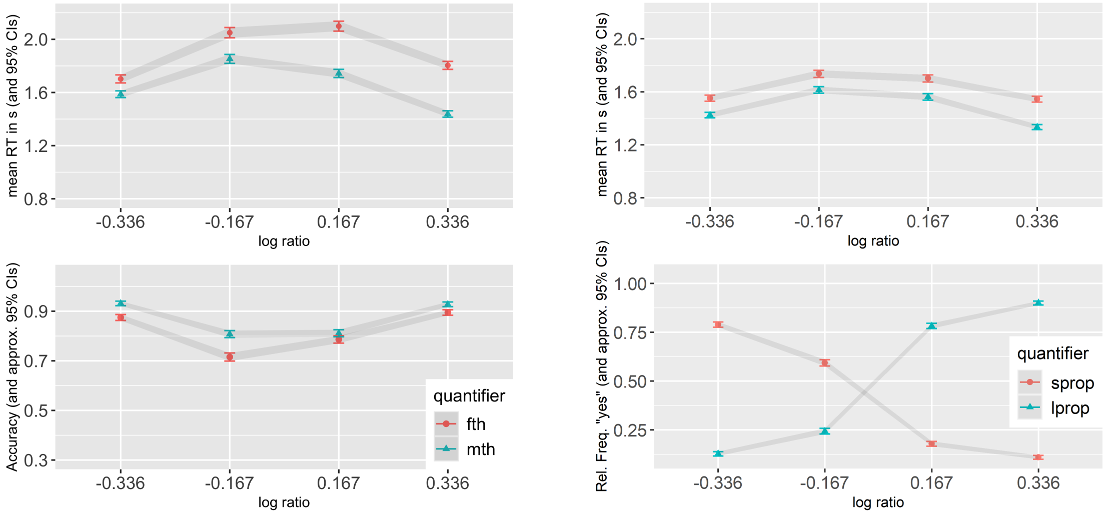
	participants					hyper means					hyper sds		
par.	dist.	p_1	p_2	l	u	dist.	p_1	p_2	l	u	dist.	p_1	p_2
t_0	tnorm	0.44	0.08	0	4	tnorm	0.44	0.04	0.1	3	gamma	1	1
$\text{logit}(z)$	tnorm	0	0.05	-2	2	tnorm	0	0.02	-1.5	1.5	gamma	1	1
$\log r_0$	tnorm	0	0.5	-1	1	beta	1	1	-0.8	0.8	gamma	1	1
V_l	tnorm	-2.5	1	-6	0	tnorm	-2.5	0.5	-5	0	gamma	1	1
g	tnorm	5	2.5	0	20	beta	1	1	0	15	gamma	1	1
a	tnorm	1.8	0.4	0	8	tnorm	1.8	0.2	0	6	gamma	1	1
s_z	tnorm	0.3	0.22	0	2	tnorm	0.3	0.1	0	1.5	gamma	1	1
s_v	tnorm	1.36	0.7	0	4	tnorm	1.36	0.3	0	3	gamma	1	1

Table 2: Specification of prior distributions. p_1 and p_2 indicate location and scale of the distributions; l and u indicate the lower and upper bound respectively.

al. did not propose normal distributions, or where we used a different parameterization, in particular with $\text{logit}(z)$ and all parameters in $v(\log r)$ (cf. equation 1), we chose prior distributions closely similar to the distributions proposed by Tran et al. (2021). On the level of hyper parameters, hyper means of midpoint and growth rates of $v(\log r)$ had uniform prior distributions due to lacking prior information and the hyper means of the remaining parameters had normal priors with the same location parameters as on the participant level but with lower scale parameters. All hyper standard deviations had $\Gamma(1, 1)$ as their prior distribution.

3 Results

We first report the regression analyses of Exp. 1 and Exp. 2, showing that the central findings of of previous studies were replicated. After that, we report the scalinDDM analysis that sheds light on the decision processes that brought these findings about.



(a) Descriptive results of Exp. 1

(b) Descriptive results of Exp. 2

Figure 4: Descriptive results of Experiments 1 & 2. fth: Fewer than half; mth: More than half; sprop: A small proportion; lprop: a large proportion.

3.1 Descriptive results and regression analysis

Mean RTs and responses across all conditions are shown in Fig. 4a for Exp. 1 and Fig. 4b for Exp. 2. For the adjectives, relative frequencies of "yes"-responses are shown instead of accuracy. As explained above in section 2 (Methods), the reason is that accuracy cannot be computed for adjectives due to their vagueness. We focus first on RT in both experiments and analyze accuracy in Exp. 1 below. While *quantifiers* took overall longer to evaluate than *adjectives*, we observed a polarity effect of about 200 ms in RT of both expression types (*adjectives*: 1485 ms vs. 1631 ms; *quantifiers*: 1655 ms vs. 1913 ms), consistent with previous studies. In addition and also consistent with previous results, larger absolute log ratios led to shorter RT (1540 ms vs. 1781 ms, resp.). Since we were first and foremost interested in the interaction between polarity and the expression type, we provide separate interaction plots showing interactions in log-RT for all four combinations of absolute log ratios and appropriateness in Fig. 5.

The combined regression analysis of RTs in Exps. 1 & 2 revealed a three-way interaction between the factors polarity, expression type and absolute log ratio ($\beta = -.0280, t = -2.15, p = .031$) as well as between the factors polarity, expression type and appropriateness ($\beta = -.0376, t = -2.89, p = .003$). To resolve these interactions and test for the critical interaction between polarity and expression type in both absolute log ratios, we analyzed the two absolute log ratios separately. Results of these analyses are shown in Table 3.

With smaller absolute log ratios, the interaction between polarity and expression type was significant ($\beta = -0.0609, t = -4.10, p < .001$) irrespective of ap-

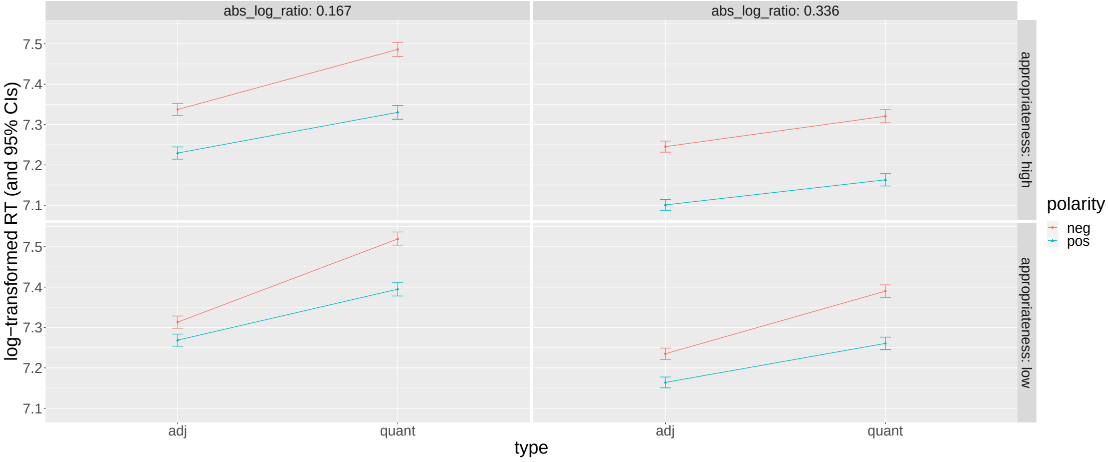


Figure 5: Log-transformed RTs in all conditions of Exps. 1 & 2 show the expected non-cross-over interaction. adj: Adjective (Exp. 2); quant: Quantifier (Exp1 1); neg: negative polarity; pos: positive polarity; abs_log_ratio: the absolute log of the ratio of yellow and blue dots.

PRORIATENESS, but APPRORIATENESS interacted with both POLARITY and EXPRESSION TYPE. To resolve the latter interactions, we computed separate analyses for lowly and highly appropriate conditions and found effects of POLARITY and EXPRESSION TYPE in both cases. Crucially, we found the predicted non-crossover interaction between EXPRESSION TYPE and POLARITY in both subsets (lowly appropriate conditions: $\beta = -.078, t = -5.79, p < .001$; highly appropriate conditions: $\beta = -.045, t = -3.30, p < .001$). As expected, the effect of POLARITY was larger for *quantifiers* than for *adjectives*, but still significant for both EXPRESSION TYPES (*adjectives*: $\beta = -.049, t = -5.54, p < .001$; *quantifiers*: $\beta = -.126, t = -12.26, p < .001$).

For larger ABSOLUTE LOG RATIOs the analysis revealed a three-way interaction between POLARITY, EXPRESSION TYPE and APPRORIATENESS. A separate analysis of lowly appropriate conditions revealed the same type of non-cross over interaction between POLARITY and EXPRESSION TYPE ($\beta = -.055, t = -4.457, p < .001$) as observed for the small ABSOLUTE LOG RATIOs. In this subset, the effect of POLARITY was significant in both expression types (*adjectives*: $\beta = -.049, t = -5.54, p < .001$; *quantifiers*: $\beta = -.126, t = -12.26, p < .001$). For the highly appropriate conditions there was, however, only the effect of POLARITY ($\beta = -.150, t = -24.09, p < .001$), but the interaction with EXPRESSION TYPE was not significant.

In Exp. 1, the effects of POLARITY and ABSOLUTE LOG RATIOs in RT were a accompanied by effects in accuracy with around 5.1% higher accuracy for *positive* than *negative* POLARITY (86.9% vs. 81.8%, resp.) and 12.6% higher accuracy for larger than smaller ABSOLUTE LOG RATIOS (88.3% vs. 75.7%, resp.). In negative vs. positive adjectives, fewer "yes"-responses were given overall and we observed a less clear-cut differentiation between the two response categories across log ratios. The main results of the regres-

Table 3: Results of combined regression analyses of RTs in Exps. 1 & 2. ALR stands for ABSOLUTE LOG RATIO, POL for POLARITY, APP for APPROPRIATENESS and TYP for TYPE. Theoretically relevant effects are also highlighted in the text.

:small absolute log ratios			
	β	t	p
TYP	.155	3.01	$= .003$
POL	$-.102$	-7.88	$< .001$
APP	$-.016$	-1.95	$= .051$
TYP×POL	$-.061$	-4.10	$< .001$
POL×APP	.040	4.24	$< .001$
APP×TYP	.045	4.79	$< .001$
::lowly appropriate conditions only			
TYP	.204	2.83	$= .005$
POL	$-.049$	-5.37	$< .001$
TYP×POL	$-.078$	$--5.79$	$< .001$
:::lowly appropriate adjectives only			
POL	$-.049$	-5.54	$< .001$
:::lowly appropriate quantifiers only			
POL	$-.126$	-12.26	$< .001$
::highly appropriate conditions only			
TYP	.147	2.38	$= .005$
POL	$-.109$	-11.86	$< .001$
TYP×POL	$-.045$	-3.30	$< .001$
:::highly appropriate adjectives only			
POL	-0.109	-7.73	$< .001$
:::highly appropriate quantifiers only			
POL	-0.154	-14.78	$< .001$

:large absolute log ratios			
	β	t	p
TYP	.076	1.680	$= .095$
POL	$-.1445$	-11.35	$< .001$
APP	$-.008$	-1.00	$= .319$
TYP×POL	$-.012$	-0.719	$= .473$
POL×APP	.072	6.108	$< .001$
APP×TYP	.077	6.236	$< .001$
APP×POL×TYP	$-.042$	-2.413	$= .016$
::lowly appropriate conditions only			
TYP	.153	3.295	$= .001$
POL	$-.072$	-8.749	$< .001$
TYP×POL	$-.055$	-4.457	$< .001$
:::lowly appropriate adjectives only			
POL	-0.049	-5.543	$< .001$
:::lowly appropriate quantifiers only			
POL	-0.126	-12.26	$< .001$
::highly appropriate conditions only			
POL	-0.150	-24.09	$< .001$

sion analysis of accuracy in Exp. 1 are given in Table 4. There were reliable effects of the factors POLARITY, ABSOLUTE LOG RATIO and TRUTH VALUE. The former two effects were due to increased accuracy for *positive* vs. *negative* POLARITY and larger vs. smaller ABSOLUTE LOG RATIO. The effect of TRUTH VALUE was due to lower accuracy with *true* vs. *false* TRUTH VALUE. In addition, we replicated the interaction between TRUTH VALUE and POLARITY in accuracy. This was due to the fact that the POLARITY effect was more pronounced for the *true* than for the *false* sentences. We also found that the factor ABSOLUTE LOG RATIO interacted with POLARITY in accuracy. These effects were due to more pronounced effects of POLARITY for larger ABSOLUTE LOG RATIOs. To resolve the TRUTH VALUE×POLARITY-interaction we analyzed the *true* and *false* conditions separately and thus tested for effects of POLARITY independently of TRUTH VALUE. The effect of POLARITY was significant in all cases and larger in true than in false conditions (true: $\beta = 0.55$; false: $\beta = 0.14$).

Table 4: Results of regression analysis of accuracy in Exp. 1. ALR stands for ABSOLUTE LOG RATIO, POL for POLARITY and TRU for TRUTH VALUE.

Accuracy for Quantifiers			
	β	z	p
TRU	-.44	-5.18	$< .001$
ALR	.87	12.12	$< .001$
POL	.19	3.28	.001
POL×TRU	.32	4.54	$< .001$
ALR×POL	.34	3.14	$< .001$
true conditions only			
MON	.55	9.24	$< .001$
false conditions only			
MON	.14	2.24	.025

3.1.1 Interim discussion

The regression analysis showed that Exp. 1 & 2 replicated previous results. In particular, we found POLARITY effects across the board and they were larger for *quantifiers* than for *adjectives* except in the conditions with *high* APPROPRIATENESS and *large* ABSOLUTE LOG RATIO. Moreover, we also replicated the interaction between POLARITY and APPROPRIATENESS (aka truth value in the *quantifier* conditions). The fact that we replicated these theoretically relevant effects allows us to study their cognitive sources using scalinDDM in the next subsection. One deviation from the results reported by Agmon et al. (2019) was the absence of the POLARITY × EXPRESSION TYPE interaction in the conditions with *high* APPROPRIATENESS and *large* ABSOLUTE LOG RATIO. The absence of this effect is hard to interpret because the log-transformation we used for RT, following Agmon et al. (2019), is itself a non-linear transformation that is used for technical reasons but theoretically unmotivated. To overcome ambiguities such transformations may cause in the context of removable interactions, we apply a scalinDDM in the next subsection to our data as a theoretically well -motivated cognitive model of choice RT.

3.2 scalinDDM analysis: Model fit, posterior distributions and hypothesis tests

In the previous section we saw that some of the central findings from the previous literature replicated in the present experiments. Specifically, these were effects of POLARITY and LOG RATIOS as well as an interaction between POLARITY and TRUTH VALUE. In the current section we use scalinDDM to investigate how the underlying decision processes brought about these effects.

3.2.1 Exp. 1: proportional quantifiers.

After a burn-in of 16500 samples (500 iterations, 33 chains; repeated until there were no "stuck" chains by the function `h.run.unstuck.dmc`), posterior distributions for individual parameters were estimated based on 330 samples each (100 iterations, 33 chains;

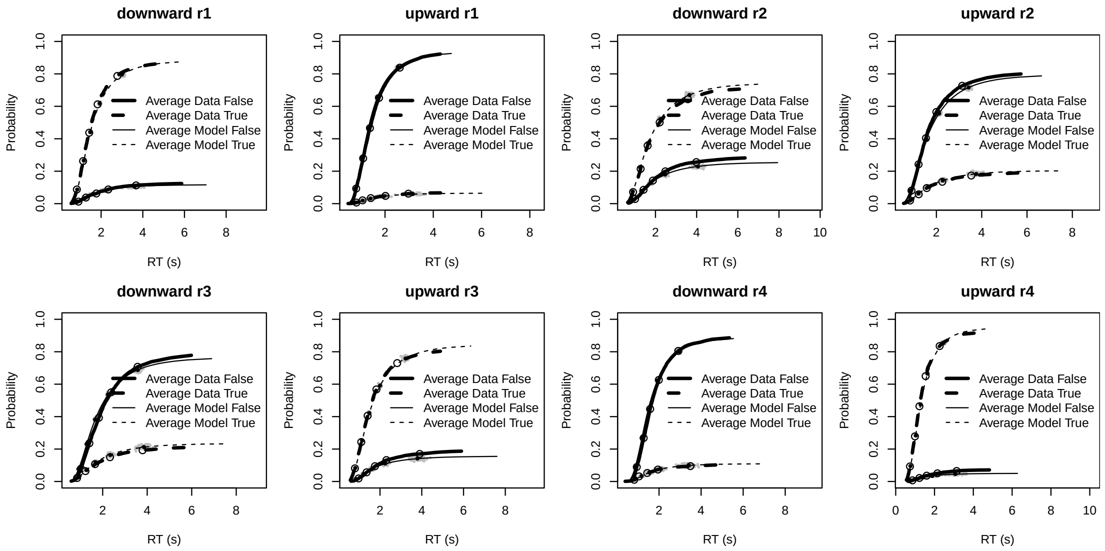


Figure 6: Defective cumulative density functions showing model fit in Exp. 1.

after discarding initial 100 samples and thinning of 10, i.e. keeping every 10th out of 3300 actual samples, and after the function `h.run.converge.dmc` automatically discarded another 200 (2 × 100) iterations because they did not help convergence). The overall model fit to the data from Exp. 1 in all eight experimental conditions was good, as shown in Figure 6 using defective cumulative density functions. Posterior distributions of hyper parameters are shown in Figure 12 in the appendix. To test for effects of the POLARITY manipulation, we examined 95% CIs of the posterior distributions of average individual parameters and of average differences between parameters (both on the level of participants)[6] in the positive and negative conditions. We found significant differences between all three parameters that were allowed to vary across POLARITY after the constraining procedure described above: non-decision time (positive: [0.592,0.604]; negative: [0.642,0.654]; difference: [0.043,0.057]; in seconds), asymptotes of the drift rate function (as a measure of differences in drift rate; positive: [2.59, 2.717]; negative: [1.892, 1.993]; difference: [0.659,0.771]); and, finally, mid points of the drift rate function (positive: [-0.026, -0.016]; negative: [-0.017, -0.003]; difference: [0.003,0.019])). Posterior distributions of these three parameters are further discussed below, where posteriors are also compared between experiments.

3.2.2 Exp. 2: adjectival expressions.

Model fit to the data from Exp. 2 was also good, as can be seen in Figure 7. As in Exp. 1 we used a burn-in of 16500 samples (500 iterations, 33 chains) and posterior distributions for individual parameters were estimated based on 39600 samples each

[6] This is the default behavior of `compare.p` and also what is recommended in DMC for the analysis of within-subject contrasts.

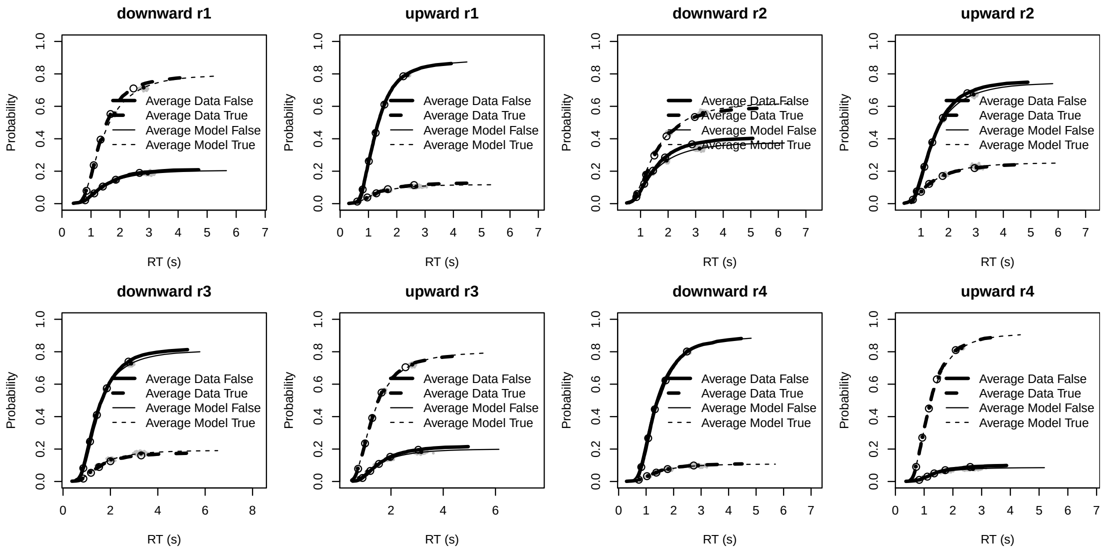


Figure 7: Defective cumulative density functions showing model fit in Exp. 2.

(1200 iterations, 33 chains; after thinning of 10 and after another 500 iterations were automatically discarded). Posterior distributions of the hyper parameters are shown in Figure 13 in the appendix. Again, we found significant differences between all three parameters that were still allowed to vary after applying parameter constraints: non-decision time (positive: [0.546,0.555]; negative: [0.603, 0.614]; difference: [0.052, 0.064]), asymptotes of the drift rate function (positive: [2.072, 2.198]; negative: [1.639, 1.741]; difference: [0.392, 0.499]); and, finally, mid points of the drift rate function (positive: [-0.029, 0.016]; negative: [-0.096, -0.079]; difference: [0.055, 0.075]).

3.2.3 Comparison between experiments.

Figure 8 shows posterior distributions of average parameters across experiments 1 and 2. As we were interested in an interaction between the TYPE of expression and its POLARITY we also looked at distributions of average differences in parameters between positive and negative expressions in the two experiments, as shown in Figure 9. As can be seen in Figure 9, differences in non-decision time between positive and negative expressions were almost perfectly constant across the two experiments. What we see in non-decision times are two main effects but no interaction: positive expressions led to shorter non-decision times than negative ones and adjectival expressions led to shorter non-decision times than proportional quantifiers. The magnitude of both of these effects was about 50 ms.

In the other two parameters, i.e. asymptotes and midpoints of the drift rate function, we see clear differences between distributions of differences in the two experiments, akin to standard interaction effects. In asymptotes, the POLARITY effect is significant in both experiments but larger in quantifiers than in adjectives: Drift rates get largest for the positive comparative quantifier, *mth*, followed in decreasing order by significantly lower

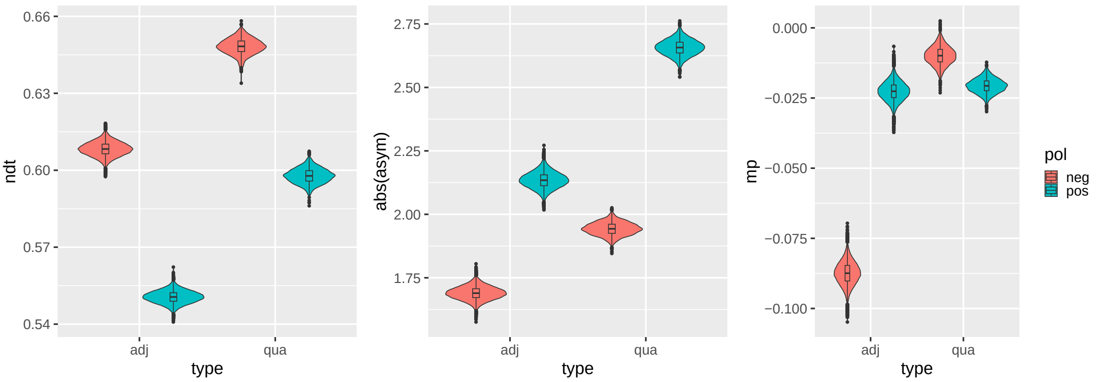


Figure 8: Posterior distributions across experiments.

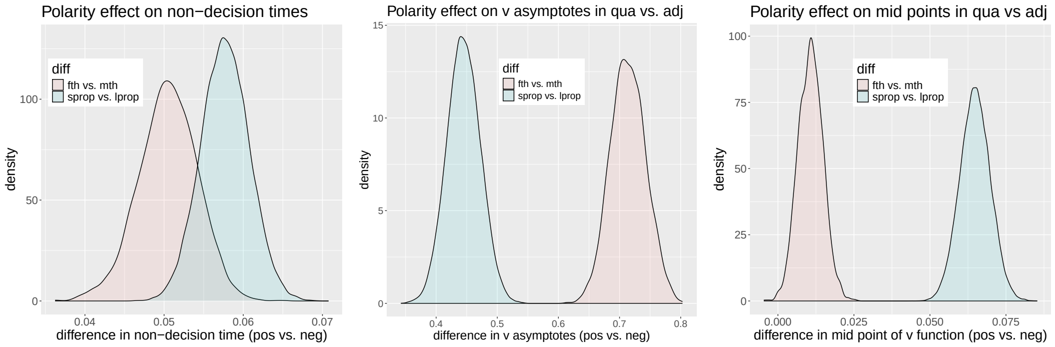


Figure 9: Posterior distributions of differences between polarities across experiments.

drift rates in the positive adjectival expression, *lprop*, the negative quantifier, *fth*, and the negative adjective, *lprop*, which do not differ much from each other.

In midpoints of the drift rate functions, the difference between positive and negative is larger for the adjectival expressions than for the proportional quantifiers. All midpoints are shifted slightly towards negative values, i.e. the switch from negative to positive drift rates occurs at proportions slightly below 0.5. Drift rates of negative adjectives had more negative midpoints than those of positive ones whereas in quantifiers the opposite was observed. Positive expressions had midpoints located close to zero with only a small difference between them. Midpoints of negative quatifiers pattern with the two postive experssions weheras midpoints of negative adjectives are shifted substantially further towards neagtive values.

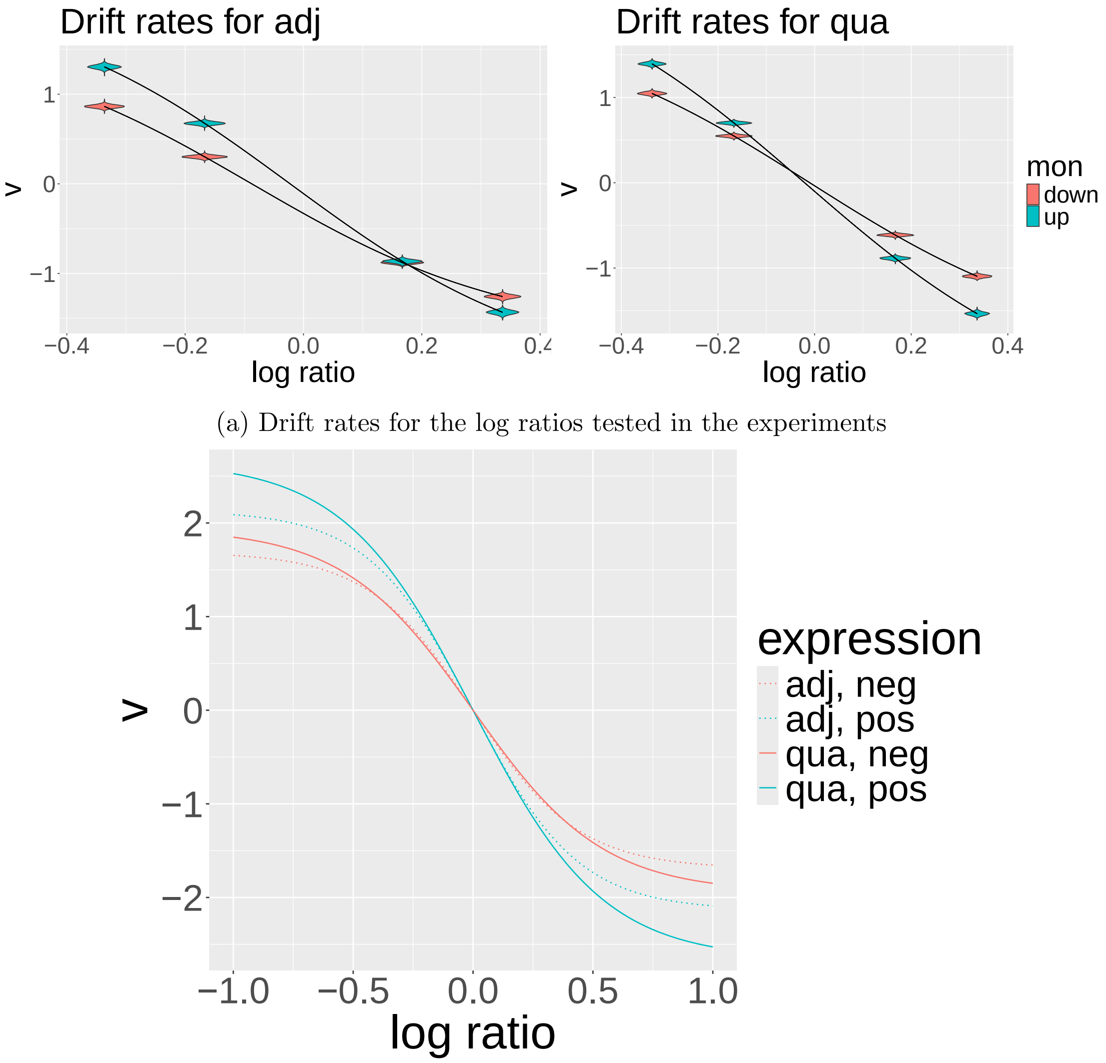
Starting points and the scale parameter of the drift rate function were estimated separately for the quantifiers and adjectives. For the quantifiers, logit-transformed starting points were positive (95% CI: [.003, 0.03], corresponding to relative starting points in the range [.501, 507]). Because of the way decision boundaries were defined in the model, this means that starting points were slightly closer to the decision boundary corresponding to the "yes, true" response for positive quantifiers and closer to the "no, false" boundary for negative quantifiers. [7] For the adjectives we observed the opposite pattern (with 95% CIs [-.039, -.012] for logit-transformed starting points corresponding to relative starting points in the range [.490, 497]).

The scale parameter of the drift-rate function was higher – indicating a faster increase from one asymptote to the other – in adjectives (95% CI: [4.336, 4.730]) than in quantifiers (95% CI: [3.570, 3.816]). This seems surprising at first given the results of Ramotowska et al. (2023), who relate this parameter to vagueness. However, in the present study asymptotes and growth rates determine the steepness of the drift rate functions together. Figure 10, specifically subfigure 10b, shows drift rate functions for both experiments. It shows qualitatively the same correspondence between vagueness and steepness of the drift function as Ramotowska et al. (2023) found. This numerical effect is, however, more pronounced in positive than in negative expressions. We come back to this issue in the discussion.

4 Discussion

In this study, we investigated how the polarity and the monotonicity of different expressions contribute to the processing cost of each of those expressions. Building on the approach of Agmon et al. (2019), we contrasted two types of quantity expressions: quantifiers (*fewer than half* vs. *more than half*) and adjectives (*the small proportion* vs. *the large proportion*). The former differ in monotonicity and polarity, and the latter in polarity only. We replicated the non-crossover interaction effect in RTs found by Agmon

[7] Starting points had different interpretations for positive and negative expressions. Positive values meant proximity to the "yes, true" decision boundary for negative expressions but proximity to the "no, false" boundary for positive ones.



(a) Drift rates for the log ratios tested in the experiments

(b) Midpoint-adjusted v functions

Figure 10: Drift rates and drift rate functions.

et al. (2019). The difference in the processing cost between quantifiers was greater than between adjectives.

As we argued in the introduction, the observation of such an interaction is insufficient to draw reliable conclusions about the underlying cognitive processes related to the linguistic properties of different expressions. This is because of multiple possible mappings between cognitive processes and observed data. Thus, based on the experimental data one cannot conclude that monotonicity adds processing cost in addition to polarity.

We employed the scalinDDM model to investigate the nature of the observed interaction and the mapping between observed effects in RTs with latent parameters of the model. In particular, we tested if the interaction in the RT data is reflected in the interaction of the magnitude of model parameters. We considered a scenario in which the interaction in RT is due to a difference in two parameters in the case of quantifiers and only one parameter in the case of adjectives. We contrasted this scenario with a scenario in which the interaction is removable and the effects in RT are additive. We also considered the possibility that the interaction would be present only for one model parameter.

Our findings turned out to be as in the last scenario. We found an interaction between expression type and polarity in parameters associated with drift rate and additive effects in non-decision time. In particular, we found a larger effect on drift rates in the quantifiers than in the adjectives and this explains a larger magnitude of the polarity effect in the former as compared to the latter type of expressions. Thus our results provide an explanation of the interaction Agmon et al. (2019) observed in RT, which we replicated here, in terms of scalinDDM parameters.

Nevertheless, the simplest interpretation of the findings reported by Agmon et al. (2019) is challenged by our modeling results because they are not consistent with an exclusive mapping between the semantic properties POLARITY and monotonicity on the one hand and scalinDDM parameters on the other. One alternative that is still in line with their conclusions would be to assume that both properties affect drift rates and this is the reason why drift rates are lower for negative than positive and, in addition, for downward entailing than upward entailing expressions. This would explain a larger difference in drift rates between *mth* and *fth* than between *lprop* and *sprop*. However, we consider this option implausible for two reasons. Firstly, the interaction in drift rates is due to the fact that *mth* has the highest drift rates among the tested conditions. This stands in contrast to the observed RT which indicate that *mth* is more difficult than *lprop*. There is thus a discrepancy between cognitive difficulty as measured in overall RT as compared to drift rates. Secondly, the difference we observed in non-decision time remains unexplained in that type of account. It would therefore call for additional assumptions in order to capture for the present results completely.

We conclude from these considerations that we are still missing a coherent and complete interpretation of the estimated model parameters at this point. Schlotterbeck et al. (2020) provided an intrepretation of their results that is worth reconsidering in this context. It is based on two influential competing theoretical accounts of the polarity effect: the two-step and the pragmatic models. We propose a modification of their interpretation below after we introduce these accounts briefly as background. Where appropriate during the following discussion, we comment on linkage assumptions that can be used to connect these two theoretical accounts with scalinDDM parameters.

4.1 Theoretical accounts of the polarity effect

According to some prominent models, the polarity effect stems from an increased complexity of negative expressions (e.g., due to hidden negation). We will call **two-step model** a broad class of models which share a common assumption that the processing of negation is carried out in a sequence of cognitive processing stages. In the case of sentence negation, the first step involves processing the positive counterpart of the negative expression and the second step involves the application of the negation (e.g. H. H. Clark, 1976; H. H. Clark & Chase, 1972; Kaup et al., 2006). For example, the sentence "The car is not red." is represented in the first step as "The car is red". Next, in the second step, the negation is applied NOT(The car is red).

The two-step model can also be extended to implicit negations. It has been argued that expressions such as negative quantifiers (e.g., *few*, *fewer than half*) contain an additional covert syntactic operator (Agmon et al., 2019; Deschamps et al., 2015; Just & Carpenter, 1971; Schlotterbeck, 2017), which increases their complexity compared to positive counterparts (e.g., *many*, *more than half*). Again, the first processing step involves the representation of a positive expression and the second step the application of the hidden negative operator (ANTONYM) as shown in Example (10).

All two-step models share the common assumption that the representation and the verification of negative expressions are carried out in steps. The longer reaction times correspond to a greater number of steps either related to a more complex representation or a longer verification procedure.

Similar to two-step models, pragmatic models also constitute a broad class, but they propose a radically different explanation of the polarity effect. In this view, processing difficulties related to negative expressions come about because negatives are systematically dispreferred compared to their positive counterparts. According to the most parsimonious explanation, negative expressions cause difficulty simply because they are used less frequently than positive ones. The less frequent an expression is, the more difficult it is to retrieve during processing (Futrell, 2024). While the simplicity of this explanation is appealing, it runs quickly into a problem of circularity. On the one hand, the least frequent expressions are the most difficult to process. On the other hand, speakers would tend to avoid expressions that are more difficult to process. As a result, a lower frequency of negatives can also be explained by processing difficulties.

To break out of this circularity, some pragmatic explanations root preferences in informativeness (e.g. Nordmeyer & Frank, 2014; Xiang et al., 2020). The informativeness-based account starts with the observation that negative expressions are often less informative than positive ones. For example, the sentence "This car is red." is, in a precise sense (Nordmeyer & Frank, 2014), more informative about the color of the car than the sentence "This car is not red". This observation is known as the principle of negative uninformativeness (Horn, 1989). While negative uninformativeness does not straightforwardly generalize from sentence negation to other forms of negatives (cf. the examples in (1), which are truth conditionally equivalent and thus also provide the same amount information about the context), we show below that it can be generalized under certain assumptions.

Some pragmatic accounts furthermore assume that low informativity may have downstream consequences that can also cause processing difficulty. An example is the dynamic pragmatic account of Tian et al. (2010). They assume that negative uninformativeness may trigger the accommodation of a Questions Under Discussion (QUD; Roberts, 2012) against which the negative expression is actually relevant and informative, e.g. *Is the car red?*. We will come back to the dynamic pragmatic account in the general discussion below.

Taken together, the pragmatic account offers a range of mechanisms explaining the polarity effect. All of these explanations refer to difficulties in the integration of negative expressions into the context. In this view, there is nothing inherently difficult to negation processing and the polarity effect arises when negation is not licensed properly by context.

4.2 Interpretation of parameters

In this section we propose an interpretation of the estimated scalinDDM parameters. In our view, this interpretation is consistent, coherent and plausible given the theoretical background laid out in the previous section. However, our interpretation does not follow deductively from the results and competing interpretations may also be viable. Below, we address some alternative interpretations and explain why we consider them implausible.

NDT and representational complexity: The two-step model explains the polarity effect in terms of a silent negation operator. Our assumption that negative quantifiers but not adjectives contain such an extra operator was derived from previous empirical results (Just & Carpenter, 1971) and also from the stance of Heim (2006, 2008) in a theoretical debate with Büring (2007, 2008) on this topic. Combining this with the linkage assumption proposed by Schlotterbeck et al. (2020), namely that representational complexity affects NDT, we expected that negativity in adjectives vs. quantifiers would affect NDT differently. The current results are, however, more in line with Büring's claim that both negative quantifiers and adjectives contain a hidden negation (cf. also Tucker, Tomaszewicz, & Wellwood, 2018, for converging evidence). In fact, if we adopt Büring's perspective, we find a close correlation between the non-decision times we estimated and the number of operators (or symbols) in the symbolic representations assumed in formal semantics (as shown in (10) below; cf. Solt, 2014). [8] NDT may thus depend on the number of operations in symbolic meaning representations and, if so, even serve as evidence in debates like the one between Heim (2006, 2008) and Büring (2007, 2008)

[8] In these representations ER is the comparative operator; POS is the so-called positive operator, which originally used to analyze positive form adjectives; ANOTONYM is a scale-reversal operator; MANY(X,Y) and PROP(X,Y) return cardinality respectively proportion of the Xs that are Y; SIZE evaluates the size of an entity (in the present case a proportion); and, finally, HALF(X) returns half the cardinality of X.

regarding hidden negation in negative adjectives. In particular, we found an increase in NDT of about 50 ms per additional operator. Negative expressions show about 50 ms longer NDT than positive ones due to the antonym operator (ANTONYM, corresponding to hidden negation) and the comparative operator (ER) in the comparative quantifiers adds an additional 50 ms. The observed differences of 50 ms extra per operator is consistent with the results of Schlotterbeck et al. (2020) and, interestingly, also plausible from the perspective of some cognitive architectures (see e.g. ACT-R, Anderson, 1990; Anderson et al., 2004)[9]. What's more, this interpretation is in line with results from concept learning (Feldman, 2000) showing that difficulty of concept learning also depends on the length of the corresponding logical representations. It is also in line with recent data by Sauerland, Meyer, and Yatsushiro (2024) indicating that during language acquisition young children have difficulty compressing complex meanings involving covert negation into a simple phonological form consisting of only one morpheme. They thus initially tend to produce an ungrammatical second morpheme in implicit negatives.

(10)

$$\begin{aligned}
\textsc{lprop}(X,Y) &:= POS(SIZE(PROP(X,Y))) && \text{(3 operators)}\\
\textsc{sprop}(X,Y) &:= POS(ANTONYM(SIZE(PROP(X,Y))) && \text{(4 operators)}\\
\textsc{mth}(X,Y) &:= POS(ER(MANY(X,Y),HALF(X))) && \text{(4 operators)}\\
\textsc{fth}(X,Y) &:= POS(ANTONYM(ER(MANY(X,Y),HALF(X))) && \text{(5 operators)}
\end{aligned}$$

Drift rates mirror a combination of frequency and informativity: Under the pragmatic model, negative expressions are difficult to process because they are generally dispreferred. This account predicts more or less directly that the polarity effect is reflected in drift rates. The reasoning behind this prediction is straightforward: Less preferred expressions are less frequent and lexical frequency has empirically been shown to modulate drift rates, e.g. in lexical decision tasks (Ratcliff & Gomez, 2004), with larger drift rates, i.e. faster evidence accumulation, for more frequent expressions. Therefore, the negative expressions that are, by hypothesis, relatively dispreferred should also lead to smaller drift rates, by analogy to other tasks. The present results confirm this prediction. Moreover, the interaction effect we found in drift rates across the two expression types (see Figures 8 and 9) qualitatively mirrors corpus frequencies of the tested expressions. In particular, we found in a corpus search that MTH is the most frequent among the tested expressions and also that the difference between positive and negative expressions is more pronounced for the quantifiers than for the adjectives.

While this interpretation of the pattern we found in drift rate estimates may seem convincing, it is faced with at least two problems. One is technical and another one is more conceptual in nature. We discuss these two problems next and hint at how they could be overcome. The technical problem is that effects of lexical frequency on drift rates in tasks like lexical decision, where words are processed out of context, need not necessarily carry over to the processing of complete sentences in visual contexts, as was

[9] 50 ms is needed for one production rule in ACT-R.

investigated in the present study (cf. the related discussion on surprisal effects during reading in Smith & Levy, 2013). In addition, it is not obvious to what degree effects of corpus frequencies may be overridden by adaptations that are caused by the repeated exposure to the expressions tested in a psycholinguistic experiment (for discussion of these issues see, e.g., van Tiel et al., 2021; Franke, 2024; Schlotterbeck, Augurzky, Ulrich, 2022)

In addition, a conceptual problem is that explanations of the polarity effect that are based purely on frequency can easily become circular, as discussed in the previous section. For example, the observed differences in frequency may simply be a consequence of differences in processing difficulty that may, in turn, be due to factors like representational complexity or monotonicity. This would mean that differences in representational complexity can explain both effects in non-decision time and drift rate. To rule out the possibility that both effects stem from the same source (e.g. difference in representational complexity) we computed correlations between non-decision time and drift rate parameters. None of the correlations was significant and all of them were below 0.1, meaning that the frequencies do not affect both parameters.

On the conceptual level, the problem of circularity could be eliminated by grounding preferences (and thus also frequencies) in informativity, as has been done also for sentence negation (Nordmeyer & Frank, 2014; Xiang et al., 2020). We think that the principle of negative uninformativeness (Horn, 1989) can be generalized to the quantity expressions studied here. Negative uninformativeness of quantity expressions emerges if we assume a specific type of QUD, something like 'How many dots are blue and how many orange?' or 'What colors do the dots have?', in a setting where there are at least three possible colors. In this kind of scenario the negative expressions are intuitively less informative about the context than the positive ones because they do not even specify the dominant color shown on the picture. Because of a general lower informativity, negative expressions are produced less often (cf. van Tiel et al., 2021). Under the assumption that participants behavior in the verification task reflects their production preferences (REF), low informativity can explain processing difficulty because uninformative expressions tend to be produced less often and, as before, infrequent expressions are in general more difficult to process than more frequent ones. If we further assume that production probabilities of utterances in context are correlated with drift rates in the verification task (cf. Schlotterbeck et al., 2020), main effects of polarity can be explained in terms of informativity.

However, we do not see how such an explanation could account for the interaction between POLARITY and EXPRESSION TYPE that we observed in drift rates. We think something extra is needed to provide an explanation of the interaction we observed and we take the interaction to indicate a separate role of frequency beside a potential effect of informativity.

Alternatively drift rates could mirror focus patterns and threshold retrieval: An alternative interpretation of differences in the drift rate parameter appeals to findings of Just and Carpenter (1971) and their hypothesis about different focus patterns of negative quantifiers and adjectives. As mentioned in the introduction, both positive and negative quantifiers focus on a larger proportion, while in case of adjectives, positive adjectives focus on a large proportion and negative on a small proportion. Additional processing difficulties for negative quantifiers as compared to adjectives arise because of a conflict between what is in focus (larger proportion) and what the sentence is about (smaller proportion). Furthermore, Just and Carpenter (1971) argued that in the sentence-picture verification task, the encoding of the picture is biased by the linguistic representation of the sentence that proceeds the picture. Thus, the negative quantifiers would bias the encoding of the picture in terms of the larger proportion and negative adjectives in terms of the smaller proportion.

When verifying the sentence participants represent the quantity extracted from the picture and derive the threshold of the linguistic expression. The drift rate could reflect this process. Higher drift rates might indicate easier threshold derivation and an easier comparison process. The threshold of quantifiers is easy to derive because there is no uncertainty about its value. The derivation of the adjectives threshold is predicted to be harder because of vagueness (cf. Ramotowska et al., 2023) which increases uncertainty about the exact threshold. This observation explains the EXPRESSION TYPE effect. Moreover, it has been proposed that the representation of a larger proportion is default. The positive expression focus on the default representation which makes it easier to compare their thresholds to the quantity representation. In contrast, the negative expressions are about the quantity that requires some additional effort to be represented. This observation explains the POLARITY effect. Finally, we hypothesize that the observed interaction in drift rate is due to the focus pattern of negative quantifiers. When verifying them participants first focus on the large proportion and then shift attention to the smaller proportion. Only then they can compare this proportion to the retrieved threshold. The attention shift and the necessity to represent both proportions requires mental efforts which results in slower drift rates.

Midpoints: Midpoints represent the point at which drift rates switch signs, i.e. the log ratio with maximal uncertainty about the judgment. For quantifiers *more than half* and *fewer than half*, their truth conditional meaning predicts that midpoints are at a log ratio of 0 (i.e. a proportion of 1:1). Adjectives, in contrast, may have different thresholds than a log ratio of 0 (cf. Ramotowska et al., 2023 for *many* and *few*). In line with previous results of Ramotowska et al. (2023), midpoints were close to 0 for quantifiers. The threshold of the positive adjective was also close to 0, while the threshold of the negative adjective was below zero. That midpoints are shifted only in negative but not positive adjectives can also be explained in terms of a tradeoff between how often an adjective can be used and how informative it is when used (cf. Lassiter & Goodman, 2017). Intuitively, the lower the threshold of negative adjectives the less often the adjective is used because it applies to fewer situations. However, more extreme thresholds increase the informativity of an adjective because they apply to more specific situation. As mentioned before, negative adjectives are less informative, thus they benefit more by having lower thresholds.

Growth rate, vagueness and uncertainty: Ramotowska et al. (2023) related the growth rate parameter (g) to vagueness, a central notion in semantics and pragmatics (see e.g. Solt, 2015c, for a recent review). They found smaller growth rates for more vague expressions (e.g. gradable adjectives) than for less vague ones (e.g. proportional quantifiers). In the current study, we observed the opposite pattern, with larger growth rates for quantifiers than adjectives. Although caution is needed in interpreting this result because it stems from a comparison between two experiments with different participants, the discrepancy still calls for an explanation. The question arises how these seemingly opposite results relate to each other. To answer this question, we note first that the results of Ramotowska et al. (2023) were obtained from a comparison between conditions in which asymptotes of the drift rate functions were constrained to be equal. In that case, growth rate corresponds directly to the steepness of the drift-rate function and thus, according to Ramotowska et al. (2023), also directly to vagueness. By contrast, asymptotes could vary across POLARITY and EXPRESSION TYPE in the current study. Therefore, as can be seen in Figure 10, steepness depends on both asymptotes and growth rates in the setting of the current study. If we adjust for the observed shifts in midpoints (mainly in the adjectives) by aligning all the drift-rate functions at a midpoint of zero (see Figure 10b), we observe that, at least numerically, the increase in drift rates is in fact less steep in adjectives than in quantifiers. This pattern is found across POLARITY but slightly more pronounced in positive than negative expressions. There is thus no direct conflict between the results of Ramotowska et al. (2023) and the present findings as long as we understand their interpretation of the growth rate in terms of overall steepness of the drift-rate function and not only in terms of the parameter g itself.

In the context of the present study, we see three factors that affect overall steepness of the drift-rate function. These are vagueness (in adjectives vs. quantifiers), uncertainty in numerosity estimation (cf. approximate vs. precise numerosity in the comparison between the visual vs. purely linguistic verification task, respectively, in Schlotterbeck et al., 2020) and POLARITY (reflected in its effect on drift rate asymptotes). The first two can be interpreted straightforwardly because both are known to affect uncertainty in the truth-value judgment task directly and this uncertainty is reflected in lower drift rates in the DDM (cf. Ramotowska et al., 2023; Schlotterbeck et al., 2020). That steepness of the drift rate function is also affected by POLARITY requires further explanation, however.

Back to the theme of pragmatics vs. representational complexity: Schlotterbeck et al. (2020) assumed a strict mapping between NDT and the complexity of semantic representations on the one hand and between drift rates and pragmatic effects on the other. In the current study, we relaxed this assumption and asked more generally which parameters are affected differently by POLARITY in the two EXPRESSION TYPES. Nevertheless a transparent interpretation of model parameters in terms of linguistic representations and processes, as proposed by Schlotterbeck et al. (2020) and Ramotowska et al. (2023), still turned out to be crucial for our present purpose, with regard to both the specific comparison between quantifiers vs. adjectives and the general evaluation of the present approach.

One aspect of the present design is useful in substantiating the interpretation of the free model parameters of scalinDDM that we developed in the previous paragraphs. According to common assumptions in semantics and pragmatics, the comparative quantifiers involve a more complex semantic representation (i.e. a more complex logical form) than the adjectives whereas the adjectives involve more complex pragmatic reasoning than the quantifiers. In particular, the comparative quantifiers encode a comparison between two cardinalities explicitly in the semantics, which leads to a more complex representation, involving more semantic operations. By contrast, the comparison is left implicit in the adjectives because we presented them in their positive form (e.g. *large*) and the threshold for comparison (e.g. the threshold at which a set of objects is considered large) is based on context-dependent pragmatic reasoning.

With this distinction in mind we see that NDT indeed corresponds to the complexity of semantic representations because quantifiers have larger NDT than adjectives. Similarly, drift rate indeed reflects the difficulty of pragmatic reasoning because the adjectives have lower drift rates than the quantifiers.

5 Conclusions

In the present study, we modeled RT and responses of two types of polar opposite expressions in a sentence-picture verification task using a theoretically driven analysis involving a generative computational model. The first pair of opposites, the proportional quantifiers, differed in monotonicity besides polarity whereas the second, involving positive-form gradable adjectives, differed only in polarity. We investigated different sources of the polarity contrast in RT and errors in these two types of expressions. In a regression analysis of RT data, we replicated the polarity effect for both types of expressions and the interaction between polarity and expression type. We further applied the computational model to test if this interaction can be removed. Our results revealed an interaction in one crucial model parameter, namely drift rates, that can account for differences between the two types of expressions. In addition, we found a second source of difficulty across negative expressions in non-decision times. We proposed an interpretation of our complete modeling results that is couched in prominent theoretical accounts of the polarity effect and of the meaning of the scalar expressions in general. Our modeling approach provides a nuanced perspective of the data and thus opens up many new avenues for future research. For example, our proposal that non-decision time is sensitive to representational complexity whereas drift rate is related to pragmatic aspects can be tested further by comparing relative adjectives (e.g. *large* vs. *small*) to absolute adjectives (e.g. *empty* vs. *full*) in their positive and comparative form. Moreover, the idea that drift rates are sensitive to frequency of use in context can be studied further by comparing expressions with identical semantic representations in different languages and different contexts.

References

Agmon, G., Bain, J. S., & Deschamps, I. (2021). Negative polarity in quantifiers evokes greater activation in language-related regions compared to negative polarity in adjectives. *Experimental Brain Research*, *239*(5), 1427–1438.

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).

Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2022). Negative Sentences Exhibit a Sustained Effect in Delayed Verification Tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *1*(48), 122–141.

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R., Bothell, D., Byrne, M., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–60. doi: 10.1037/0033-295X.111.4.1036

Augurzky, P., Schlotterbeck, F., & Ulrich, R. (2020). Most (but not all) quantifiers are interpreted immediately in visual context. *Language, Cognition and Neuroscience*, *35*(9), 1203–1222.

Bader, M., & Haeussler, J. (2010, 07). Toward a model of grammaticality judgments. *Journal of Linguistics*, *46*, 273 - 330. doi: 10.1017/S0022226709990260

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. doi: 10.1037/0033-295X.113.4.700

Bott, O., Schlotterbeck, F., & Klein, U. (2019). Empty-Set Effects in Quantifier Interpretation. *Journal of Semantics*, *36*(1), 99-163.

Brasoveanu, A., Clercq, K. D., Farkas, D., & Roelofsen, F. (2014). Question tags and sentential negativity. *Lingua*, *145*, 173–93. doi: 10.1016/j.lingua.2014.03.008

Büring, D. (2007). *More* or *less*. In *Paper presented at the chicago linguistic society meeting.* Chicago.

Büring, D. (2008). The least *at least* can do. In C. B. Chang & H. J. Haynie (Eds.), *Proceedings of the 26th west coast conference on formal linguistics* (pp. 114–120). Somerville, MA: Cascadilla Proceedings Project.

Clark, H. H. (1976). *Semantics and Comprehension.* The Hague: Mouton & Co.B.V.

Clark, H. H., & Chase, W. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472-517. doi: 10.1016/0010-0285(72)90019-9

Clark, R. G., Blanchard, W., Hui, F. K., Tian, R., & Woods, H. (2023). Dealing with complete separation and quasi-complete separation in logistic regression for linguistic data. *Research Methods in Applied Linguistics*, *2*(1), 1-11. doi: 10.1016/j.rmal.2023.100044

Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In P. Haggard, Y. Rosetti, & M. Kawato (Eds.), *Sensorimotor Foundations of Higher Cogni-*

tion (p. 527-574). Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199231447.001.0001

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology*, *122*, 371-396.

Denić, M., Homer, V., Rothschild, D., & Chemla, E. (2021). The influence of polarity items on inferential judgments. *Cognition*, *215*, 104791. doi: 10.1016/j.cognition.2021.104791

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition*, *143*, 115–28.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. doi: 10.1016/j.tics.2004.05.002

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6084), 630–633. doi: 10.1038/35036586

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry, N. W. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, *20*(4), 400–409. doi: 10.1111/j.1469-8986.1983.tb00920.x

Futrell, R. (2024). An information-theoretic account of availability effects in language production. *Topics in Cognitive Science*, *16*(1), 38-53. doi: https://doi.org/10.1111/tops.12716

Greenberg, J. H. (1963). Universals of language.

Grodzinsky, Y., Agmon, G., Snir, K., Deschamps, I., & Loewenstein, Y. (2018). processing cost of Downward Entailingness: the representation and verification of comparative constructions. *ZAS Papers in Linguistics*, *60*, 435–451.

Grodzinsky, Y., Jaichenco, V., Deschamps, I., Sánchez, M. E., Fuchs, M., Pieperhoff, P., ... Amunts, K. (2020). Negation and the Brain. In V. Dèprez & T. M. Espinal (Eds.), *The Oxford Handbook of Negation* (pp. 693–712). Oxford University Press.

Heathcote, A., Lin, Y.-S., Reynolds, A., Strickland, L., Gretton, M., & Matzke, D. (2019). Dynamic models of choice. *Behavior research methods*, *51*, 961–985.

Heim, I. (2006). Decomposing antonyms? In *Proceedings of semantics and linguistic theory conference* (Vol. 16, pp. 35–58). doi: 10.3765/salt.v16i0.2941

Heim, I. (2008). Decomposing antonyms? In *Proceedings of sinn und bedeutung* (Vol. 12, pp. 212–225).

Horn, L. R. (1989). *A natural history of negation.* The University of Chicago Press.

Huang, Y., & Ferreira, F. (2020). The application of signal detection theory to acceptability judgments. *Frontiers in Psychology*, *11*. doi: 10.3389/fpsyg.2020.00073

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

Kang, I., & Ratcliff, R. (2020). Modeling the interaction of numerosity and perceptual variables with the diffusion model. *Cognitive Psychology*, *120*, 101288.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*(4), 498–525. doi: 10.2307/1418556

Kaup, B., & Dudschig, C. (2020). Understanding Negation: Issues in the processing of negation. In *The Oxford Handbook of Negation.* Oxford University Press. doi: 10.1093/oxfordhb/9780198830528.013.33

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, *38*(7), 1033–1050.

Knowlton, T., Pietroski, P., Halberda, J., & Lidz, J. (2022). The mental representation of universal quantifiers. *Linguistics and Philosophy*, *45*, 911–941. doi: 10.1007/s10988-021-09337-8

Lassiter, D., & Goodman, N. (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, *194*, 3801–36. doi: 10.1007/s11229-015-0786-1

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of 'most'. *Natural Language Semantics*, *19*(3), 227–256. doi: 10.1007/s11050-010-9062-6

Loftus, G. (1978). On interpretation of interactions. *Memory& Cognition*, *6*, 312–319. doi: 10.3758/BF03197461

Moxey, L. M., & Sanford, A. J. (1986). Quantifiers and focus. *Journal of semantics*, *5*(3), 189–206.

Moxey, L. M., Sanford, A. J., & Dawydiak, E. J. (2001). Denials as controllers of negative quantifier focus. *Journal of Memory and Language*, *44*(3), 427–442.

Mulder, M., van Maanen, L., & Forstmann, B. (2014). Perceptual decision neurosciences – a model-based review. *Neuroscience*, *277*, 872–84. doi: 10.1016/j.neuroscience.2014.07.031

Nicenboim, B., Schad, D. J., & Vasishth, S. (2021). *Introduction to bayesian data analysis for cognitive science.* Under contract with Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series.

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 the 36th annual conference of the cognitive science society.*

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 1-1. Retrieved from `https://doi.org/10.1167/5.5.1` doi: 10.1167/5.5.1

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of 'most': Semantics, numerosity, and psychology. *Mind and Language*, *24*(5), 554–585. doi: 10.1111/j.1468-0017.2009.01374.x

Potthoff, R., Ramotowska, S., Szymanik, J., & van Maanen, L. (2023). Time-pressure does not alter the bias towards canonical interpretation of quantifiers. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Ramotowska, S., Steinert-Threlkeld, S., Van Maanen, L., & Szymanik, J. (2020). Most, but not more than half, is proportion-dependent and sensitive to individual differences. In M. Franke, N. Kompa, M. Liu, J. L. Mueller, & J. Schwab (Eds.), *Proceedings of sinn und bedeutung* (Vol. 24, pp. 165–182). doi: 10.18148/sub/2020.v24i2.891

Ramotowska, S., Steinert-Threlkeld, S., van Maanen, L., & Szymanik, J. (2023). Uncovering the structure of semantic representations using a computational model of decision-making. *Cognitive Science*, *47*(1), e13234.

Ratcliff, R., & Gomez, G., P. amd McKoon. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*(1), 159–82. doi: 10.1037/0033-295X.111.1.159

Ratcliff, R., & McKoon, G. (2018). Modeling numerosity representation with an integrated diffusion model. *Psychological Review*, *125*(2), 183–217. doi: 10.1037/rev0000085

Ratcliff, R., & McKoon, G. (2020). Examining aging and numerosity using an integrated diffusion model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(11), 2128.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in cognitive sciences*, *20*(4), 260–281.

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, *5*, 6–1.

Rück, F., Dudschig, C., Mackenzie, I., Vogt, A., Leuthold, H., & Kaup, B. (2021). The role of predictability during negation processing in truth-value judgment tasks. *Journal of Psycholinguistic Research*, *50*, 1437–1459. doi: 10.1007/s10936-021-09804-0

Sauerland, U., Meyer, M.-C., & Yatsushiro, K. (2024). The cum-sine pattern in german child language: An argument for antonym decomposition. *Language Acquisition*, *0*(0), 1–13. doi: 10.1080/10489223.2024.2332452

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). doi: 10.15496/publikation-18745

Schlotterbeck, F., Ramotowska, S., Van Maanen, L., & Szymanik, J. (2020). Representational complexity and pragmatics cause the monotonicity effect. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. ((Decision expected on May 1st))

Schoeller, A., & Franke, M. (2015). Semantic values as latent parameters: Surprising few & many. In S. D'Antonio, M. Moroney, & C.-R. Little (Eds.), *Semantics and linguistic theory (SALT)* (Vol. 25, p. 143-162). LSA and CLC Publications. doi: 10.3765/salt.v25i0.3058

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–19. doi: 10.1016/j.cognition.2013.02.013

Solt, S. (2014, 01). Q-Adjectives and the Semantics of Quantity. *Journal of Semantics*, *32*(2), 221-273. doi: 10.1093/jos/fft018

Solt, S. (2015a). Measurement scales in natural language. *Language and Linguistics Compass*, *9*(1), 14-32. doi: https://doi.org/10.1111/lnc3.12101

Solt, S. (2015b). Vagueness and imprecision: Empirical foundations. *Annual Review of Linguistics*, *1*, 107–127. doi: 10.1146/annurev-linguist-030514-125150

Solt, S. (2015c). Vagueness and imprecision: Empirical foundations. *Annual Review of Linguistics*, *1*(1), 107–27. doi: 10.1146/annurev-linguist-030514-125150

Tian, Y., Breheny, R., & Ferguson, H. J. (2010). Why we simulate negated information: A dynamic pragmatic account. *Quarterly Journal of Experimental Psychology*, *63*(12), 2305–2312.

Tian, Y., Ferguson, H., & Breheny, R. (2016). Processing negation without context–why and when we represent the positive argument. *Language, Cognition and Neuroscience*, *31*(5), 683–698.

Tran, N.-H., van Maanen, L., Heathcote, A., & Matzke, D. (2021). Systematic parameter reviews in cognitive modeling: Towards a robust and cumulative characterization of psychological processes in the diffusion decision model. *Frontiers in Psychology*, *11*, 3922. doi: 10.3389/fpsyg.2020.608287

Tucker, D., Tomaszewicz, B., & Wellwood, A. (2018). Decomposition and processing of negative adjectival comparatives. In E. Castroviejo, L. McNally, & G. Weidman Sassoon (Eds.), *The semantics of gradability, vagueness, and scale structure: Experimental perspectives* (pp. 243–273). Cham: Springer International Publishing. doi: 10.1007/978-3-319-77791-7_10

Urbach, T. P., DeLong, K. A., & Kutas, M. (2015). Quantifiers are incrementally interpreted in context, more than less. *Journal of Memory and Language*, *83*, 79–96. doi: 10.1016/j.jml.2015.03.010

Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, *63*(2), 158–179. doi: 10.1016/j.jml.2010.03.008

Vanek, N., Matic Škoric, A., Košutar, S., Matějka, u., & Stone, K. (2024). Mental simulation of the factual and the illusory in negation processing: evidence from anticipatory eye movements on a blank screen. *Scientific Reports*, *14*, 2844. doi: 10.1038/s41598-024-53353-0

van Tiel, B., Franke, M., & Sauerland, U. (2021). Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences*, *118*(9). doi: 10.1073/pnas.2005453118

van Tiel, B., & Pankratz, E. (2021). Adjectival polarity and the processing of scalar inferences. *Glossa: a journal of general linguistics*, *6*(1), 32. doi: 10.5334/gjgl.1457

Wagenmakers, E. J., Krypotos, A. M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: a survey of the field 33 years after loftus. *Cognition*, *215*, 104791. doi: 10.1016/j.cognition.2021.104791

Wason, P. C. (1961). Response to Affirmative and Negative Binary Statements. *British Journal of Psychology*, *52*(2), 133–142.

Xiang, M., Kramer, A., & Nordmeyer, A. E. (2020). An informativity-based account of negation complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(10), 1857.

Zuanazzi, A., Ripollés, P., Lin, W. M., Gwilliams, L., King, J.-R., & Poeppel, D. (2024, 05). Negation mitigates rather than inverts the neural representations of adjectives. *PLOS Biology*, *22*(5), 1-33. doi: 10.1371/journal.pbio.3002622

A Prior and posterior distributions

The following Figure 11 shows prior distributions used to fit scalinDDM model to Exps 1& 2. Figure 12 shows posterior distributions for Exp. 1 and Figure 13 shows posterior distributions for Exp. 2.

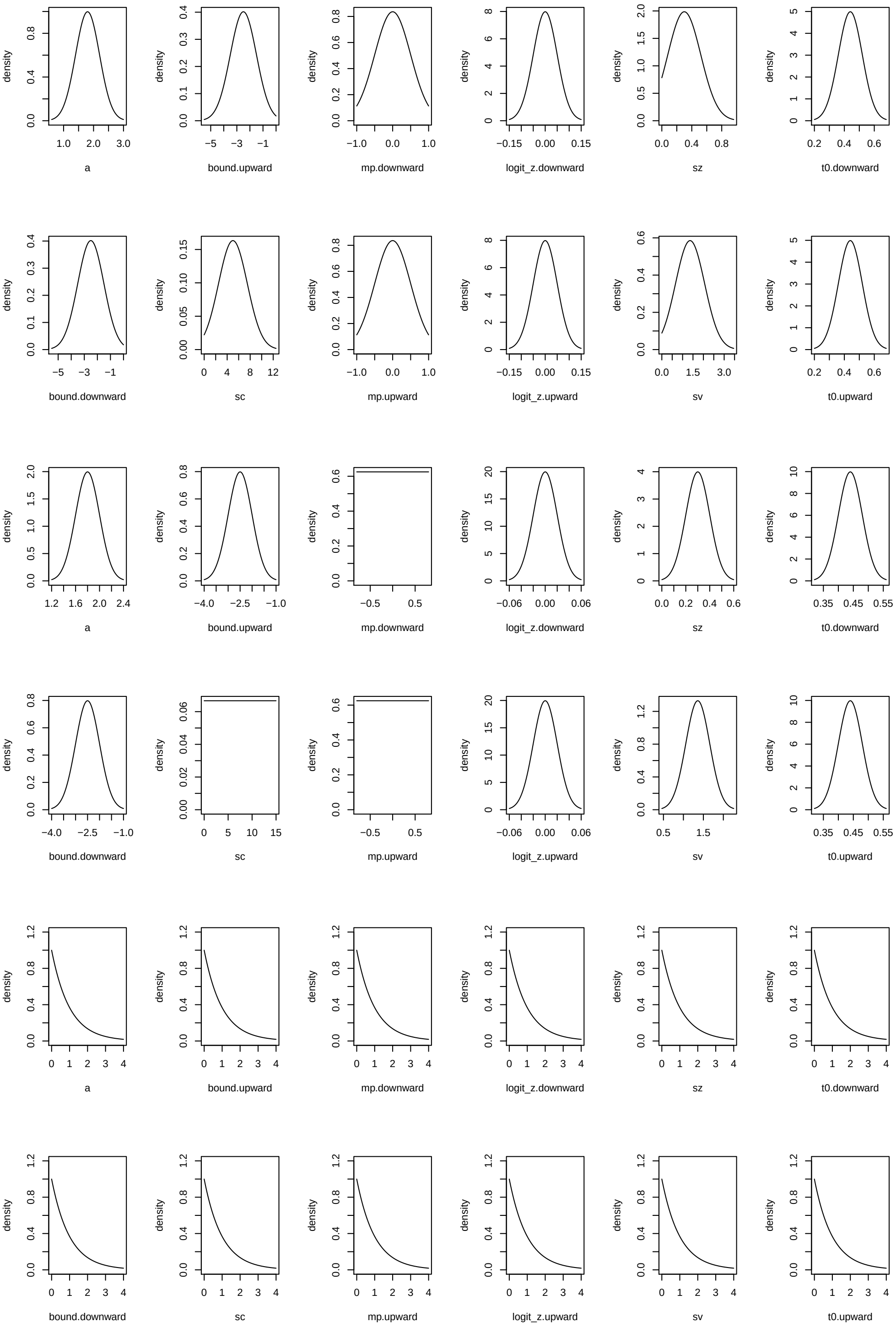


Figure 11: Priors for the 12 free model parameters: top: participant level priors; middle: priors for hyper means; bottom: priors for hyper standard deviations. [Growth rate, g, is labeled “sc” and lower asymptotes, V_l are labeled “bound”.]

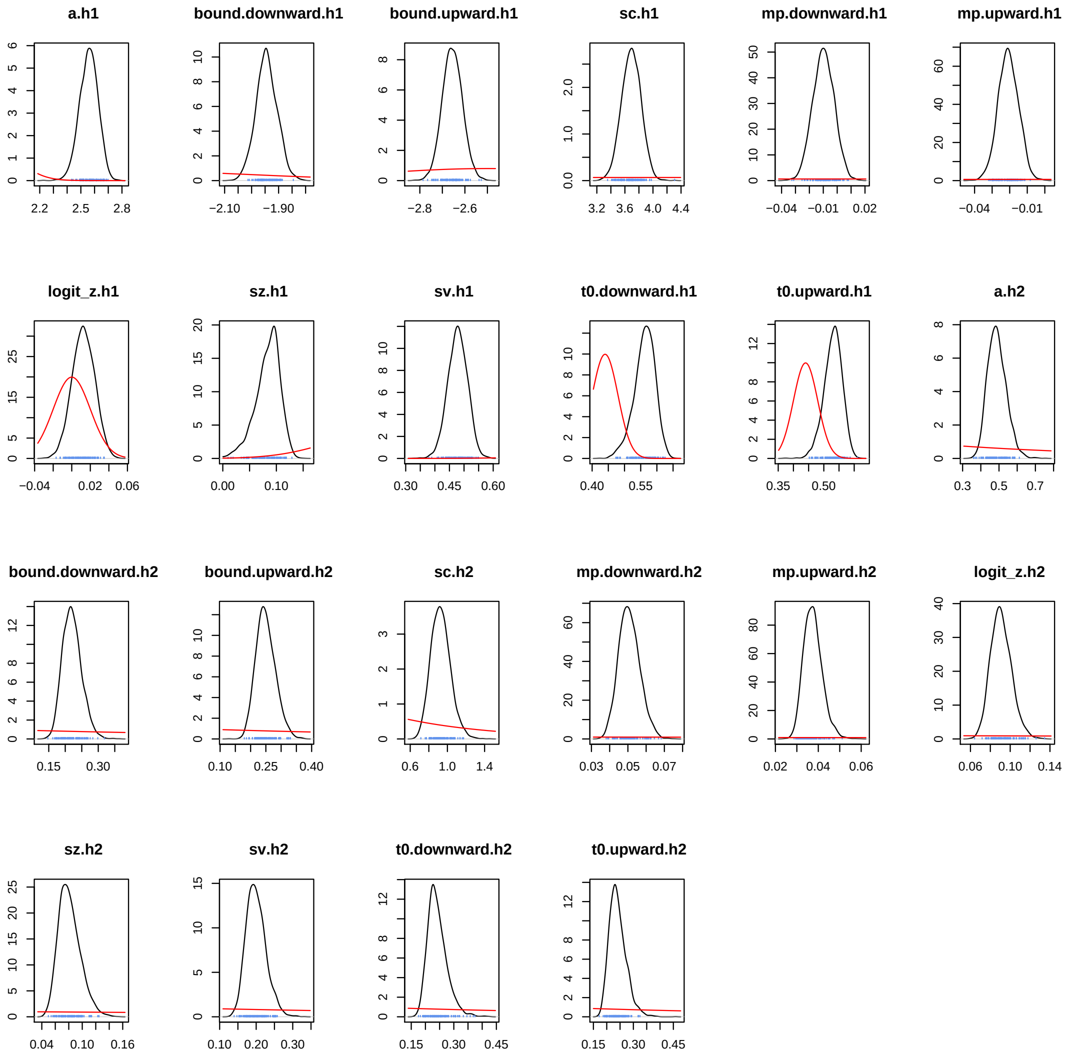


Figure 12: Posterior distributions over hyper parameters in Exp. 1 (black lines) alongside their prior distributions (red lines) and 95% CIs (blue region). [Growth rate, g, is labeled "sc" and lower asymptotes, V_l are labeled "bound".]

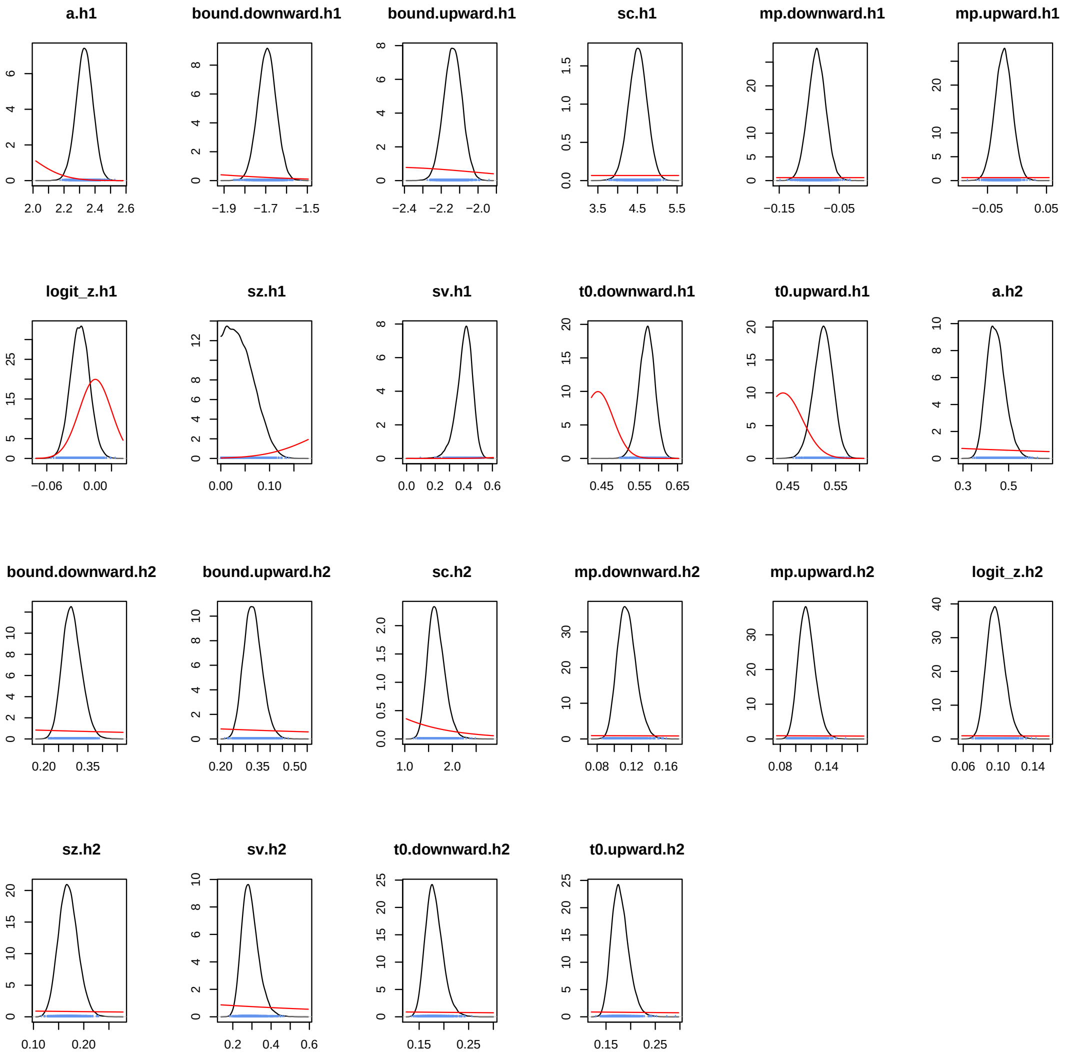


Figure 13: Posterior distributions over hyper parameters in Exp. 2 (black lines) alongside their prior distributions (red lines) and 95% CIs (blue region). [At the moment, growth rate, g, is labeled “sc” and lower asymptotes, V_l are labeled “bound”.]