Re-Examining the Scalar Polarity Hypothesis

Word count: 9,873 words

Abstract

Studies by van Tiel et al. (2019) and van Tiel & Pankratz (2021) suggest that the cognitive effort and, hence, time course associated with scalar implicature (SI) processing depends on the polarity of the underlying scale: According to this *polarity hypothesis*, SIs induced by a positively polar scale (e.g., ⟨some, all⟩) take more time to process than a literal interpretation, whereas SIs induced by a negatively polar scale (e.g., ⟨not all, none⟩) are processed faster than a literal interpretation, relative to respective baseline response times (RTs) on pragmatically unambiguous sentences. After conducting a planned experiment through online crowd-sourcing ($N = 400$), employing a sentence–picture verification task, we do not find supporting evidence for the polarity hypothesis. Rather, effect patterns vary widely across four structurally similar, positive–negative scale pairs which we assessed. This finding contributes to a more nuanced understanding of how scalar implicatures are processed. It suggests that scalar diversity (van Tiel et al. 2016) may play a more pervasive role than previously assumed, with heuristics such as polarity not being sufficient to account for diversity between the processing signatures of logically otherwise equivalent scales.

Keywords: scalar implicature; polarity; sentence processing; experimental pragmatics

1 Introduction

The following sentence,

(1) Some of the bottles are full.

has two possible interpretations: It can mean either that some *and possibly all* of the bottles in question are full (literal interpretation), or that some, *but not all* of them are (pragmatic interpretation). The apparent ambiguity of the quantifier 'some', here, is often explained through pragmatic reasoning: If a cooperative speaker (see Grice 1975) utters (1), then they are likely to have good reason to consider a semantically compatible, but more restrictive alternative to (1), i.e.,

(2) All of the bottles are full.

to be false—otherwise they should have uttered (2) instead of (1) in order to be maximally informative to their listener. Therefore, guided by the assumption that the speaker is, in fact, cooperative, certain listeners will be inclined towards a pragmatic interpretation of (1) (some, but not all …). Yet, other listeners may not hold that assumption or have a general tendency to focus on purely logical meaning, thus being inclined towards a literal interpretation of (1) (some and possibly all …).

The crucial difference between (1) and (2) lies in the replacement of 'some' with 'all'. Horn (1972) was the first to conceptualise expressions which can be replaced by each other in order to increase or decrease semantic restrictiveness as components of an ordered *scale*. In the example case, a scale of quantifier expressions ⟨some, all⟩ can be assumed: ‘some’ may be referred to as the weaker term and ‘all’ as the stronger term of that scale.

The reasoning that leads to endorsing a pragmatic interpretation of sentences like (1) is often called *scalar implicature* (alternatively, scalar inference). Scalar implicature (SI) is a phenomenon that has long been of interest to experimental pragmaticists (see, e.g., studies by Rips 1975, Noveck 2001, Bott & Noveck 2004, De Neys & Schaeken 2007, Degen & Tanenhaus 2015). Typical experimental designs aimed at detecting the cognitive computation of SIs require participants to judge the truthfulness or felicity of a sentence like (1), in circumstances where either world knowledge or complementary visual information renders the sentence true or felicitous only under a literal, but not under a pragmatic interpretation. This line of research also seems to suggest that the computation of pragmatic interpretations is associated with higher cognitive effort (see, e.g., Bott & Noveck 2004) than the computation of literal interpretations.

The present work puts a recently proposed hypothesis, which was first expressed in detail by van Tiel et al. (2019), to the test. According to that hypothesis, the cognitive effort associated with the processing of SIs crucially depends on the *polarity* of the underlying scalar term: For example, relatively high effort is predicted for processing the *positively polar* scalar term ‘some’ as implicating ‘but not all’, yet relatively low effort for processing the *negatively polar* scalar term ‘not all’ as implicating ‘but not none’ (= ‘but some’).

After carrying out a properly powered experiment which consisted of a sentence–picture verification task (see Section 2.1), administered through web-based crowdsourcing, we do not find supporting evidence for the polarity hypothesis. This null finding may prompt a re-evaluation of that particular account. It could guide future work to a deeper exploration of potential factors modulating the cognitive effort associated with SI processing. Such an endeavour would effectively extend work by van Tiel et al. (2016) who delineate other scale properties (such as availability, distinctness, etc.) which might be responsible for observed between-scale processing differences (scalar diversity).

1.1 Defining Scalar Polarity

Several criteria can be employed to decide whether the polarity of a scale $\langle\alpha,\beta\rangle$ is positive or negative: A useful heuristic to start with is to find a scale $\langle\neg\beta,\neg\alpha\rangle$ which is antonymous to $\langle\alpha,\beta\rangle$, i.e., where the new scale’s weaker term $\neg\beta$ can be interpreted as a negation of the original scale’s stronger term β and the new scale’s stronger term $\neg\alpha$ as a negation of the original scale’s weaker term α. In the case of the quantifier scale ⟨some, all⟩, it is easy to see how its antonym scale can be formed and then linguistically simplified, as shown below in i.:

i. $\langle\alpha,\beta\rangle = \langle\text{some, all}\rangle \implies \langle\neg\beta,\neg\alpha\rangle = \langle\textbf{not}\text{ all, }\textbf{not}\text{ some}\rangle = \langle\text{not all, none}\rangle$

ii. $\langle\gamma,\delta\rangle = \langle\text{not all, none}\rangle \implies \langle\neg\delta,\neg\gamma\rangle = \langle\textbf{not}\text{ none, }\textbf{not}\text{ not all}\rangle = \langle\text{some, all}\rangle$

Likewise, as shown in ii., the original scale ⟨some, all⟩ can be derived as the antonym of the scale ⟨not all, none⟩. The bidirectional transformability of these scales into antonyms of each other, showcased here, demonstrates that they are polar opposites.

In order to establish which one of two such opposing scales is positively polar, it is useful to think about whether the terms that constitute a scale introduce lower bounds on the meaning dimension that they describe (van Tiel et al. 2019). For instance, this is the case for the terms ‘some’ and ‘all’ as they introduce the constraints $> 0\,\%$ and $= 100\,\%$, respectively, on whatever quantitative dimension they refer to. Therefore, the scale ⟨some, all⟩ can be considered positively polar. To complement this, a negatively polar scale's terms typically introduce upper bounds on their meaning dimension. Indeed, 'not all' and 'none' introduce the constraints $< 100\%$ and $= 0\%$ on any quantity. Hence, the scale ⟨not all, none⟩ is negatively polar.

Note that the same logic expressed by Horn (1972) equally applies to negatively polar scales. For example, the sentence (4), using the stronger scalar term 'none', is a semantically compatible, but more informative alternative to (3), which uses the weaker scalar term 'not all':

(3) Not all of the bottles are full.

(4) None of the bottles are full.

However, this semantically informed definition of scalar polarity is not the only one that has been proposed. In related research, polarity is sometimes more broadly understood as capturing *both* such negation semantics *and* the further-reaching concept of emotional valence (e.g., van Tiel & Pankratz 2021): Under this view, both ⟨not all, none⟩ (negation semantics) and ⟨drizzly, rainy⟩ (negative emotional valence) are considered legitimate negative scales. Moreover, van Tiel & Pankratz (2021) maintain a further distinction between scales like ⟨not all, none⟩ and scales like ⟨drizzly, rainy⟩: The former is an *explicitly* negative scale since it contains explicit morphemic markers of negativity ({not}, {none}). The latter, by contrast, is an *implicitly* negative scale as no such explicit negation markers are found in 'drizzly' or in 'rainy'. This distinction, first laid out by Fodor & Garrett (1975), is worth mentioning because van Tiel and Pankratz lay out slightly differential predictions for implicitly vs. explicitly negative scales: The effect in verification times postulated by the polarity hypothesis should be even stronger, they argue, for explicitly negative scales than for implicitly negative ones. In their own study, they only test implicitly negative scales—with the notable exception of ⟨unlikely, impossible⟩ which does contain the negation morphemes {un-} and {im-}.

1.2 Scalar Implicature Processing

In a seminal study, Bott & Noveck (2004) conducted several experiments showing that, with regard to a French equivalent of the ⟨some, all⟩ scale, participants responded significantly more slowly when judging the veracity of underinformative sentences pragmatically (as 'False') rather than literally (as 'True'), but did not show an analogous difference in response times between 'False' and 'True' verification on unambiguous sentences used as control trials. An example of an underinformative sentence tested in the study by Bott and Noveck is given in (5).

(5) Certains éléphants sont des mammifères.
some elephants are IND-PL mammals
'Some elephants are mammals.'

This finding came to be known as the *B&N effect*. For the same scale, De Neys & Schaeken (2007) found that participants who had to perform a dual task, consisting of both sentence verification and memorisation of complex dot patterns, ended up verifying underinformative sentences literally (rather than pragmatically) more often than control subjects. Those control subjects performed only the verification task and whose cognitive capacity was thus not additionally charged during their language comprehension process. This latter result is sometimes referred to as the *D&S effect*. Both the B&N and the D&S effect can be seen as consistent with Relevance theory (Sperber & Wilson 1986), but incompatible with the neo-Gricean view (Levinson 2000) of scalar implicature processing. That is, the effects suggest that drawing SIs is additionally effortful compared to interpreting underinformative sentences literally.

1.3 The Polarity Hypothesis

As already briefly mentioned, rather recently, a further theory has been put forward (van Tiel et al. 2019; van Tiel & Pankratz 2021), claiming that the B&N effect—and by extension also the D&S effect—is present only for positively polar scales, like ⟨some, all⟩. By contrast, negatively polar scales, like ⟨not all, none⟩, should show a *reversed* B&N effect (i.e., literal responses take longer than pragmatic ones) or, in some cases, at least the absence of any effect.

Apart from their own studies' findings, van Tiel and colleagues point towards additional evidence that can be interpreted, partially, in support of this polarity hypothesis (evidence from: Cremers & Chemla 2014; Romoli & Schwarz 2015; Marty et al. 2020). Their theoretical explanation for why pragmatic interpretations arising from negative scales are computed equally as fast as or even faster than corresponding literal ones (unlike what is the case for positive scales) goes as follows: They refer back to a model of the time course of negation processing, proposed by Clark & Chase (1972), which argues that positively polar scalars (e.g., 'above') are processed very quickly, whereas (explicitly) negatively polar ones (e.g., 'not above') are associated with higher cognitive and, thus, temporal processing cost. Based on this distinction, van Tiel & Pankratz (2021) theorise that a positively polar surface utterance evokes a negatively polar implicature once interpreted pragmatically. This is shown in (6)—where '⇝' abbreviates 'implicates' and '+'/'−' positive/negative polarity—thus delaying processing due to the effort associated with cognitively focusing on negation (⇒ B&N effect). By contrast, for a negatively polar surface utterance, cognitively focusing on the literal meaning rather than on the implicature, as shown in (7), the latter being positively polar here, is more effortful (⇒ absent or reversed B&N effect).

(6) [Some]$_+$ of the bottles are full. ⇝ [Not all]$_-$ (…)

(7) [Not all]$_-$ of the bottles are full. ⇝ [Some]$_+$ (…)

Van Tiel et al. (2019) found evidence consistent with the polarity hypothesis after conducting several web-based experiments that employed a sentence–picture verification paradigm: Their results showed a pattern where negatively polar scalars ('low', 'scarce') were not associated with B&N or D&S effects, while positively polar scalars ('or', 'might', 'some', 'most', 'try') mostly were. Van Tiel and Pankratz (2021) followed up on that finding with another web-based experiment, now systematically comparing sentence–picture verification responses to adjectival scalars. Again, it was found that B&N effects tended to occur only with positively polar scalars.

1.4 Limitations of Previous Work

Although the polarity hypothesis by van Tiel and colleagues offers an interesting and intuitively plausible account of SI processing, the evidence base in support of it is not that large and decisive yet.

The seminal study by van Tiel et al. (2019) had set out to experimentally compare the properties of seven different scales. Two of them, ⟨low, empty⟩ and ⟨scarce, absent⟩, happened to be the only negatively polar scales involved. Incidentally, however, these two scales were also the only adjectival scales that were included—the remaining five consisted of verbs, pronouns, or conjunctions. Only because ⟨low, empty⟩ and ⟨scarce, absent⟩ ended up standing out in comparison to the remaining scales—i.e., they generally tended not to show a B&N or D&S effect across the three performed experiments—van Tiel and colleagues were led to formulate what is now the polarity hypothesis as an exploratory claim. Clearly, this initial, exploratory (and potentially confounded) finding alone does not provide very strong support for the polarity hypothesis.

By contrast, the follow-up study by van Tiel & Pankratz (2021) can be viewed as stronger evidence for the polarity hypothesis since its very purpose was to test this hypothesis in a confirmatory manner. Importantly, it also solves the issue of a potential confound due to part-of-speech category as it exclusively compares adjectival scales amongst one another. But it displays other properties that may be viewed as limitations: First, its experimental design *does not* make direct comparisons within polarity-contrastive pairs of scales whose elements tap the same meaning dimension (e.g., comparing positive ⟨warm, hot⟩ directly against negative ⟨cool, cold⟩ on the meaning dimension of temperature or comparing positive ⟨some, all⟩ directly against negative ⟨not all, none⟩ on the meaning dimension of quantity). Rather, van Tiel and Pankratz examine an arbitrary set of six positive adjectival scales (among them, e.g., ⟨content, happy⟩, ⟨warm, hot⟩, ⟨ajar, open⟩) against an arbitrary set of ten negative adjectival scales (among them, e.g., ⟨mediocre, bad⟩, ⟨drizzly, rainy⟩, ⟨cool, cold⟩). In consequence, the reported results on processing differences between positive and negative scales during sentence–picture verification have not been controlled for potential confounding effects of particular lexical properties (as well as particularities of the respective associated sentences and pictures) of the scales that were chosen to represent either positive or negative polarity. Second, rather than treating polarity as a binary factor as is more typical, they employ a continuous latent metric based on a range of very diverse corpus-based heuristics, like frequency, appearance in 'how' questions, and human-annotated emotional valence, to rate the polarity of scales. They motivate this operationalisation choice by arguing that both the linguistic notion of 'polarity' (defined based on lower/upper bounds on intuitive meaning dimensions as discussed earlier) and the psychological notion of 'polarity' (determined by the emotional valence evoked by words, e.g., 'happy' as positive and 'sad' as negative) should be combined into a single measure. This, however, leaves it unclear what the constructed latent measure actually represents, especially considering that a linguistically positively 'polar' word (e.g., 'angry' on the meaning dimension of anger; positive since one can be twice as angry as someone else) can be seen as psychologically negatively 'polar' at the same time (to be 'angry' is typically perceived as an emotion of negative valence).

Coming back to the distinction between explicit negation (e.g., in expressions '**not** above', '**not** many', '**in**frequently', '**un**abundant') versus implicit negation (e.g., in respective synonyms 'below', 'few', 'rarely', 'scarce'),[1] mentioned in Section 1.1, another observation needs to be made: Van Tiel and Pankratz (2021) theorise that scales based on explicitly negative scalar expressions should show even a *reversed* B&N effect (i.e., literal responses take longer than pragmatic ones), whereas scales based on implicitly negative scalar expressions should just show the *absence* of any effect (i.e., literal responses take roughly the same time as pragmatic ones). The reasoning behind this hypothesis is that there may be a hierarchy of negation processing in language, with explicitly marked negation being harder to process than implicit negation. This idea is supported by classical findings of differences in sentence verification times (Clark & Chase 1972) depending on

[1] The distinction made by Fodor & Garrett (1975) classifies negation in a linguistic expression as explicitly negative if it is expressed by a corresponding morpheme like {not}, {un-}, or {im-} whose sole purpose is to convey negative meaning, but as implicitly negative if the negative information is only a partial, built-in semantic aspect of a morpheme with a more complex meaning, e.g., {low}, {few}, {rare}, or {scarce}.

the type of negation involved. It follows that the interaction between response and polarity on response times which is predicted by the polarity hypothesis should be of even greater magnitude when explicit negation is considered, as opposed to what is the case for implicit negation. However, neither van Tiel et al. (2019) nor van Tiel & Pankratz (2021) actually experimentally examine explicitly negative scales—they only test implicitly negative ones (granted, ⟨**un**likely, **im**possible⟩ in the 2021 paper can be viewed as an exception, but crucially it does not show a reversed B&N effect as would have been expected). Thus, based only on the experiments presented in the two mentioned papers by van Tiel and colleagues, the even stronger polarity-hypothesis claim regarding explicitly negative scalars remains a speculative one as it has not been systematically assessed. Nonetheless, earlier work by other authors—some of which van Tiel and colleagues also refer back to in making their case—has, indeed, compared the processing of positive SIs with that of *explicitly* negative ones. We provide an overview of that earlier work in the following two paragraphs:

The earliest study that van Tiel & Pankratz (2021) cite in support of their polarity-based view is the one by Cremers & Chemla (2014): In Cremers and Chelma's 'Experiment 1', participants were asked to verify underinformative sentences (intermixed within pragmatically unambiguous control sentences) that may provoke SI processing based on either the positive scale ⟨some, all⟩ or the negative one ⟨not all, no⟩[2] against their world knowledge. For instance, participants would be shown an underinformative sentence like 'Some elephants are mammals' and then asked to make a 'True'/'False' decision. The pattern of results from that experiment, regarding residual response times (residualised against the control trials), does, in fact, resemble what van Tiel and colleagues' polarity hypothesis would have predicted: Pragmatic responses take significantly longer than literal ones for underinformative sentences with 'some' (B&N effect), but, as opposed to that, it is literal responses that take significantly longer than pragmatic ones for underinformative sentences with 'not all' (reversed B&N effect). Interestingly, however, Cremers and Chelma themselves are quite critical in evaluating that result: They argue that it is likely caused by a confound due to particularities of the design of their control trials which were used as fillers and for residualisation. Therefore, in the very same paper (Cremers & Chemla 2014), the authors go on to present a follow-up experiment ('Experiment 2') whose design differs from 'Experiment 1' in that it (a) eliminates the potential confound in the controls, (b) uses a larger post-exclusion sample size of 60 rather than 36 subjects, and (c) relies on a paradigm where participants are divided beforehand into a literal vs. a pragmatic condition within which they are respectively instructed and trained what kind of response they should apply on target trials rather than relying on intuitive judgements.[3] Crucially, results from that 'Experiment 2' show the interaction of interest, predicted by the polarity hypothesis, to be practically zero: Here, both positive and negative SIs display a regular B&N effect.

Romoli & Schwarz (2015) conducted a similar verification experiment as Cremers & Chemla (2014) did, and they are also cited by van Tiel & Pankratz (2021) in support of the polarity hypothesis. As an important difference, however, while they did test negatively polar SIs, they did not compare them to positively polar SIs, but rather to pre-

[2] Note that what we call 'positive' / 'positively polar' vs. 'negative' / 'negatively polar' here, based on terminology used in van Tiel and colleagues' line of research, is referred to by Cremers & Chemla (2014) as 'direct (SIs)' vs. 'indirect (SIs)' instead, following a terminological and conceptual convention introduced by Chierchia et al. (2004).

[3] This instruction-based (rather than intuition-based) experimental paradigm for assessing on-line SI processing was first used by Bott & Noveck (2004) in their 'Exp. 1' and has since been replicated in several subsequent studies, including by van Tiel et al. (2019) in their 'Exp. 3'.

suppositions (a somewhat different phenomenon), in terms of respective response-time latencies. Hence, although it does partially speak in favour of the polarity hypothesis that Romoli and Schwarz' results did end up showing a *reversed* B&N effect for negatively polar SIs, their experiment lacks a comparative condition where positively polar SIs would also be tested (for which then a regular, non-reversed B&N effect would be expected).

Marty et al. (2020) present results of an experiment very close in design to the sentence–picture verification task previously employed by van Tiel et al. (2019). Yet, the design of the included linguistic materials differs from van Tiel and colleagues in that it actually ensures to compare positive scales directly against negative scales in a pair-wise manner (i.e., ⟨some, all⟩ vs. ⟨not all, none⟩; ⟨$_{NP_1}$ or $_{NP_2}$, $_{NP_1}$ and $_{NP_2}$⟩ vs. ⟨not both $_{NP_1}$ and $_{NP_2}$, neither $_{NP_1}$ nor $_{NP_2}$⟩;[4] ⟨possible, certain⟩ vs. ⟨not certain, impossible⟩) and that all three included negative scales display explicit rather than implicit negation. The results of Marty et al.'s experiment clearly show the effect pattern in response-time latencies that is assumed by the polarity hypothesis across all three positive–negative scale comparisons. In that, their study can be considered as the most salient evidence so far in favour of van Tiel and colleagues' polarity hypothesis about response-time latencies. As a minor caveat, though, Marty et al.'s experiment featured two additional, between-subject conditions where the verification task of interest was intermixed either with a low-memory-load pattern memorisation task or a high-memory-load pattern memorisation task. This was an attempt to replicate De Neys & Schaeken's (2007) experimental design and, thus, to detect potential (reversed) D&S effects. Unlike what van Tiel and colleagues may have predicted in analogy to the B&N-related polarity hypothesis, that is, here, a D&S effect for positive SIs, but a reversed D&S effect for negative SIs, in the actual results both positive and negative SIs end up displaying regular, non-reversed D&S effects across all scale pairs. In consequence, Marty et al.'s results are evidence for the polarity hypothesis only in the narrower sense concerning response-time latencies and (reversed-)B&N effect patterns. In any case, this narrower sense is what our present work is focused on as well. Lastly, a study by Bill et al. (2018) can also be interpreted through the lens of the polarity hypothesis: Their experiment compared the processing of a positively polar underinformative statement (i.e., given a picture indicating the idea of always going to the movies, the sentence: 'John sometimes went to the movies.') against that of a negatively polar one (i.e., given a picture indicating never going to the movies, the sentence: 'John didn't always go to the movies.') within a Covered Box (Huang et al. 2013) paradigm. Here, the interaction of interest to the polarity hypothesis regarding response-time latencies was essentially zero. In this case, both the positive and the negative condition display the absence of any (reversed) B&N effect.

In summary so far, we do not find much decisive support for the polarity hypothesis (regarding response-time latencies) in previous work, with the notable exception of the study by Marty et al. (2020) where the expected effect pattern clearly shows up. Further, the 2019 and 2021 studies conducted by van Tiel and colleagues in particular—based on which these authors have originally formulated the polarity hypothesis—neither directly examine polarity-contrastive scale pairs, nor do they systematically assess cases of explicit negation. Thus, in the present study, we seek to address these two points by conducting a sentence–picture verification experiment that compares the implicature processing of explicitly negative scales directly to that of their precise positive counterparts.

[4] Here, the subscripted expressions NP_1 and NP_2 are placeholders for two arbitrary noun phrases.

1.5 The Present Study

Our present study aims to collect direct evidence for the scalar polarity hypothesis insofar as it is concerned with differences in response-time patterns (B&N effects vs. absent/reversed B&N effects). In order to do so, a sentence–picture verification task was administered. During that task, participants had to judge the felicity of sentences (as 'Good' or 'Bad') with regard to subsequently displayed pictures. Cases where the sentence was an underinformative description of the picture and, therefore, pragmatically ambiguous, made up the critical trials. By contrast, in control trials, the sentence–picture combination did not constitute any pragmatic ambiguity.

The prediction that can be derived from the scalar polarity hypothesis and which we want to test here is the following:

- SIs of negative polarity (e.g., scale ⟨not all, none⟩) are processed faster than a literal interpretation, whereas SIs of positive polarity (e.g., scale ⟨some, all⟩) are processed more slowly than a literal interpretation, relative to respective baseline response times (RTs) on pragmatically unambiguous sentences.

2 Experiment

In what follows, we summarise the design and the data-collection method of our experiment. Further below (Section 2.4), the method of statistical analysis is outlined, followed by a prospective power analysis based on that method and previously collected pilot data (Section 2.5).

The methodological decisions presented here had been preregistered in a time-stamped preregistration protocol (see under Data Availability).

2.1 Materials and Design

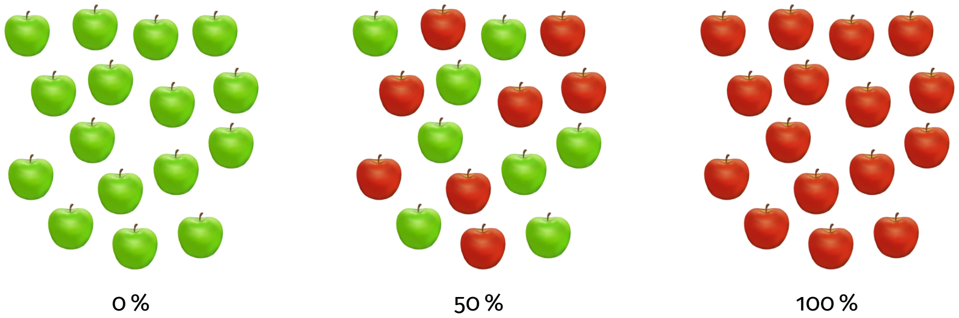


Figure 1: The three picture variants of the Quantity item, corresponding to scales ⟨some, all⟩ and ⟨not all, none⟩. The percentages annotated below indicate the degree of informative fulfilment with respect to the positive scale.

Sentence–picture verification (SPV) tasks were originally devised by Clark & Chase (1972). They have since been used in multiple studies on SIs in particular (Tavano & Kaiser 2010, Marty et al. 2013, van Tiel et al. 2019, van Tiel & Pankratz 2021, to name a few). Eight different scales are used in the present experiment, embedded within the sentences listed in Table 1. We may label and denote these eight scales as +quantity ⟨some, all⟩, −quantity

Item	Polarity	Scalar	Sentence
Quantity	+	$\exists$	**Some** of the apples are red.
	+	$\forall$	**All** of the apples are red.
	–	$\neg\forall$	**Not all** of the apples are red.
	–	$\neg\exists$	**None** of the apples are red.
Possibility	+	$\exists$	The bead **might be** falling into a blue bin.
	+	$\forall$	The bead **is definitely** falling into a blue bin.
	–	$\neg\forall$	The bead **might not be** falling into a blue bin.
	–	$\neg\exists$	The bead **is definitely not** falling into a blue bin.
Time	+	$\exists$	She **sometimes** hit the bullseye today.
	+	$\forall$	She **always** hit the bullseye today.
	–	$\neg\forall$	She ***did*** **not always** hit the bullseye today.
	–	$\neg\exists$	She **never** hit the bullseye today.
Space	+	$\exists$	**Somewhere** in Africa ***it is*** daytime.
	+	$\forall$	**Everywhere** in Africa ***it is*** daytime.
	–	$\neg\forall$	**Not everywhere** in Africa ***is it*** daytime.
	–	$\neg\exists$	**Nowhere** in Africa ***is it*** daytime.

Table 1: The sentences used in the sentence–picture verification task in all their possible variants. Within each item, each sentence variant employs a different scalar, whose polarity can be either positive (+) or negative (–). Weak positive scalars behave like an existential quantifier $\exists$ (e.g., 'some'), while strong positive scalars behave like a universal quantifier $\forall$ (e.g., 'all'). Likewise, weak negative scalars behave like a negated universal $\neg\forall$ (e.g., 'not all'), while strong negative scalars behave like a negated existential $\neg\exists$ (e.g., 'none'). Scalar terms within the sentences are marked in bold; any surrounding material that is also subject to change between sentence variants is highlighted in bold and italics.

Polarity	Scalar	Picture Variant		
		0 %	50 %	100 %
+	$\exists$	**B** $_{t3+}$	**G** $_{t2+}$	**G/B** $_{t1+}$
+	$\forall$	**B** $_{t6+}$	**B** $_{t5+}$	**G** $_{t4+}$
–	$\neg\forall$	**G/B** $_{t1-}$	**G** $_{t2-}$	**B** $_{t3-}$
–	$\neg\exists$	**G** $_{t4-}$	**B** $_{t5-}$	**B** $_{t6-}$

Table 2: Trial types *t1+*, ..., *t6+* and *t1–*, ..., *t6–* resulting from combinations of scalar-specific sentence variants ($\exists$/$\forall$/$\neg\forall$/$\neg\exists$) and picture variants (0/50/100 %), marked with corresponding type(s) of reasonable verification responses (**G** = 'Good'; **B** = 'Bad'). Trial types *t1+* (positively polar) and *t1–* (negatively polar) are the critical ones of interest as their underinformativeness reasonably allows for both 'Good' (literal) and 'Bad' (pragmatic) responses.

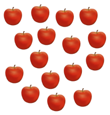
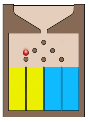
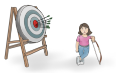
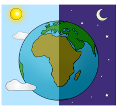
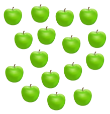
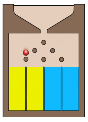
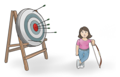
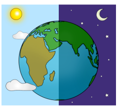
Trial ID	Sentence	Picture	Condition	Expected
$t1+q$	Some of the apples are red.		critical	**G/B**
$t1-p$	The bead might not be falling into a blue bin.		critical	**G/B**
$t1+t$	She sometimes hit the bullseye today.		critical	**G/B**
$t2-s$	Somewhere in Africa it is daytime.		control	**G**
$t3+q$	Some of the apples are red.		control	**B**
$t4-p$	The bead is definitely not falling into a blue bin.		control	**G**
$t5+t$	She always hit the bullseye today.		control	**B**
$t6-s$	Nowhere in Africa is it daytime.		control	**B**

Table 3: Eight examples of trials, of which three are critical and five belong to the control condition. The symbol '+' or '–' in a trial ID indicates the polarity of the featured scalar. The last letter in a trial ID (e.g., 'q' in $t1+q$) distinguishes between the four sentence–picture items *quantity*, *possibility*, *time*, and *space*. Overall, there are 48 (= 6 × 2 × 4) possible trials. In the rightmost column, expected verification responses (**G** = 'Good'; **B** = 'Bad') are listed.

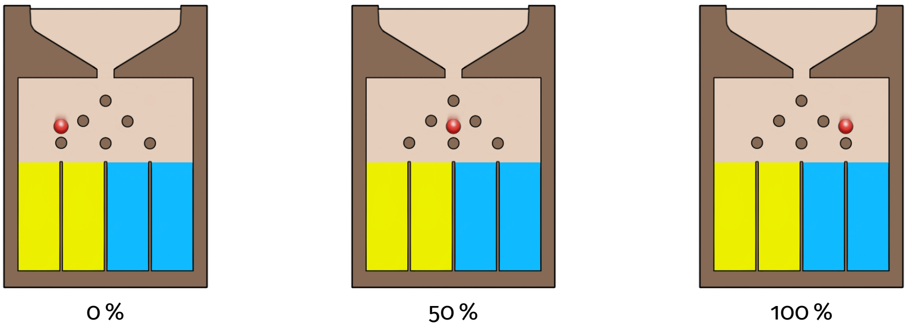


Figure 2: The three picture variants of the Possibility item, corresponding to scales ⟨might be, is definitely⟩ and ⟨might not be, is definitely not⟩. The percentages annotated below indicate the degree of informative fulfilment with respect to the positive scale.

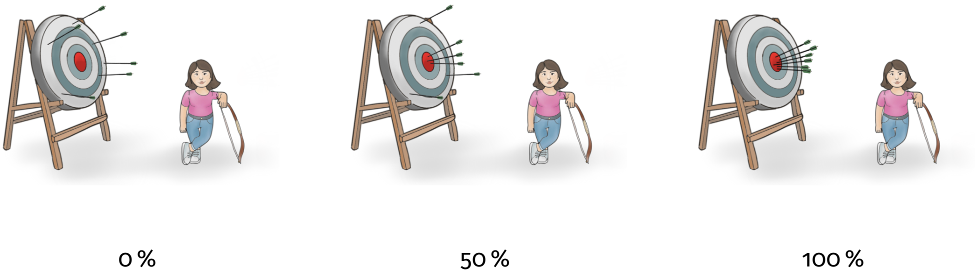


Figure 3: The three picture variants for the Time item, corresponding to scales ⟨sometimes, always⟩ and ⟨not always, never⟩. The percentages annotated below indicate the degree of informative fulfilment with respect to the positive scale.

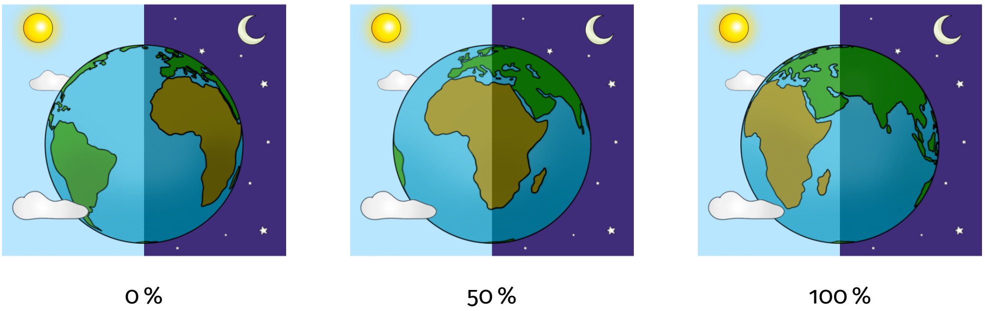


Figure 4: The three picture variants of the Space item, corresponding to scales ⟨somewhere, everywhere⟩ and ⟨not everywhere, nowhere⟩. The percentages annotated below indicate the degree of informative fulfilment with respect to the positive scale.

⟨not all, none⟩, +possibility ⟨might be, is definitely⟩, −possibility ⟨might not be, is definitely not⟩, +time ⟨sometimes, always⟩, −time ⟨not always, never⟩, +space ⟨somewhere, everywhere⟩, and −space ⟨not everywhere, nowhere⟩.

Four of these scales display positive polarity, and each positively polar scale can be contrasted with exactly one negatively polar counterpart. Hereinafter, we may refer to such polarity-contrastive pairs of scales as *items*. We label the four items as follows: Quantity, Possibility, Time, and Space (see Table 1) and abbreviate the individual scales, of which there are two in each item, as +quantity, −quantity, ..., −space, with the sign at the start of each label designating polarity.

One thing that the scales we included all have in common is that they exhibit a *logical norm of correctness*.[5] That is, as opposed to what holds for other kinds of scales like ⟨low, empty⟩ or ⟨warm, hot⟩, the present scalars' truth-value interpretation is always objectively defined in *non*-underinformative sentences as it follows the pattern of logical quantifiers in ⟨∃, ∀⟩ (positive polarity) and ⟨¬∀, ¬∃⟩ (negative polarity), respectively. To give two examples: The scalars 'some' and 'might be' behave like existential quantifiers (∃) on the respective meaning dimensions of quantity (of something measureable) and possibility (of some event). Correspondingly, 'all' and 'is definitely' behave like universal quantifiers (∀) on the same meaning dimensions. The weaker scalemates of the analogous negatively polar scales, 'not all' and 'might not be',[6] behave like negated universal quantifiers (¬∀), while their stronger scalemates, 'none' and 'is definitely not',[7] behave like negated existentials (¬∃).

In the SPV task that is administered here, these scalar expressions appear within the sentences given in Table 1. These sentences, in turn, are associated with the pictures displayed in Figures 1, 2, 3, and 4. Each item features three different picture variants, labeled 0 %, 50 %, and 100 %, respectively, where the percentage indicates the ratio of visual elements contributing to maximising the informative fulfilment of the corresponding positive scale.

In Table 2, all twelve possible combinations of types of scalars and picture variants are listed. Each such combination may be called a *trial type*. Note that the trial type [∃, 100 %], abbreviated with the label *t1*+, and the trial type [¬∀, 0 %], abbreviated with the label *t1*−, are the two critical ones of interest, here, since they are ambiguous in the sense that the underlying combinations can sensibly be interpreted either as 'Good' (literally) or as 'Bad' (pragmatically). The remaining ten trial types are labelled *t2*...*t6*± (for details, again, Table 2) and constitute pragmatically unambiguous combinations. With every trial type being featured in four different sentence–picture items, i.e., the four items labelled Quantity, Possibility, Time, and Space, this gives rise to 48 trials, of which 8 are critical ones and the remaining 40 can serve as control trials. For illustration purposes, several concrete examples of trials are listed in Table 3.

When administering this SPV task to a subject, each subject would see all 48 trials in an individually randomised order, with the following restriction being imposed on the randomisation process: Two trials from the same item would never appear immediately one after the other.

[5] Terminology borrowed from Dieussaert et al. (2011).

[6] To be very precise, the negative scale of the Possibility item, ⟨might not be, is definitely not⟩, is actually somewhat special compared to the three remaining ones in that its direct translation into a formal semantic representation would *not* yield something like $\langle \neg\forall p.Event(p), \neg\exists p.Event(p)\rangle$, but rather something like $\langle \exists p.\neg Event(p), \forall p.\neg Event(p)\rangle$. Yet, the latter kind of representation is logically equivalent to the former in terms of truth values. Therefore, we still include ⟨might not be, is definitely not⟩ as the negative counterpart to the positive scale ⟨might be, is definitely⟩, here.

[7] See footnote 6.

2.2 Procedure

The experiment was run via Prolific (**www.prolific.co**), relying on an implementation in the Lingoturk (Pusse et al. 2016) framework. Participants accessed the experiment using a common web browser on a laptop or desktop PC.

It took roughly six minutes for a participant to complete the experiment. Each participant was paid 1.06 GBP.

The sentence–picture verification task encompasses 48 different trials, each of which consists of (1) a slide reading ‘Press C on your keyboard to continue’, (2) a slide showing a sentence, replaced upon pressing the space bar by (3) a slide showing a corresponding picture, which disappears once either of the two keys **1** (‘Good’) or **0** (‘Bad’) is pressed. On every trial, the verification response and response time is registered, but participants do not receive any feedback regarding the responses they provide. The experiment ends once all trials have been presented.

2.3 Participant Sample and Exclusion Criteria

Native English speakers who are US citizens, live in the US, fall into the age range of 30–40 years (i.e., an 11-year time span) and have normal or corrected-to-normal vision were recruited as participants. (The sample size was determined based on a prospective power analysis; see Section 2.5.)

Following the practice by van Tiel & Pankratz (2021), participants who responded erroneously to more than 20 % of control trials in the SPV task were excluded from any subsequent analyses. Further, sentence–picture verification responses with response times that were either shorter than 200 ms or at least as long as 10,000 ms were removed. This was done in order to avoid treating accidental button presses (< 200 ms case) or implausibly slow responses (≥ 10,000 ms case) as valid observations that merit analysis.

2.4 Statistical Models

First, we divide the experimental response data into a subset of critical trials and a subset of control trials. Then, we fit the following Bayesian linear hierarchical model on the log-transformed RTs (below: `log_RT`) in the control data set using the *brms* package (Bürkner 2017) for R (R Core Team 2023), with `VR` denoting the verification response (sum-coded as ‘Good’ = +0.41 and ‘Bad’ = −0.59) and `POL` the polarity (sum-coded as positive = +0.5 and negative = −0.5):

Control model, formula given in Lmer syntax (data = control trials):

```
log_RT ~ VR * POL + (VR * POL | subject) + (VR * POL | item)
```

The control model uses the priors Normal(7, 1) for log RT, Normal(0, 0.3) for population-level standard deviation, Normal(0, 0.2) for any population-level main effects or interactions, Normal(0, 0.1) for any group-level standard deviations, and LKJ(2) for any correlation parameters (see Lewandowski et al. 2009).

In order to analyse response-time variation that is unique to critical trials, we residualise the log-RT variable as follows: First, we generate predicted values from the fitted control model for each combination of verification response, polarity, participant, and item. In the separate data set of critical trials, these control-predicted values are then subtracted from corresponding observed log RTs, thus giving rise to the desired new variable, *residual log RT*, for each critical-trial observation. As stated, this residualised variable has the advantage of only capturing variation that is unique to critical trials and, thus, pragmatic ambiguity—eliminating the possibility that it might be confounded by such effects of verification response or polarity that trivially also occur on control trials. For instance, there may be a trivial baseline difference between the time it takes to respond ‘Good’ versus the time to respond ‘Bad’ which is also present on unambiguous sentences and thus bears no theoretical relevance for the comparison of literal (‘Good’) and pragmatic (‘Bad’) responses on critical trials.

Next, we fit our actual model of interest on the critical-trial data set. Here, the dependent measure is residual log RT (below: `res_log_RT`):

Model of interest, formula given in Lmer syntax (data = critical trials):

```
res_log_RT ~ VR * POL + (VR * POL | subject) + (VR * POL | item)
```

The model of interest uses the priors Normal$(0, 1)$ for residual log RT, Normal$(0, 0.3)$ for population-level standard deviation, Normal$(0, 0.2)$ for any population-level main effects or interactions, Normal$(0, 0.1)$ for any group-level standard deviations, and LKJ(2) for any correlation parameters (see Lewandowski et al. 2009).

If the polarity hypothesis is true, then we can indeed expect a `VR:POL` interaction effect, with a negative-sign parameter estimate $\hat{\beta}_{\texttt{VR:POL}}$. We assess the posterior estimate of the interaction parameter by computing $P(\beta_{\texttt{VR:POL}} > 0)$, i.e., the posterior probability that the parameter value is positive. If we find $P(\beta_{\texttt{VR:POL}} > 0) < 0.05$, then we may consider the result to be consistent with the polarity hypothesis. Another metric we report is the Bayes factor in favour of the polarity hypothesis. We compute the Bayes factor BF_{10} using the Savage–Dickey method (Verdinelli & Wasserman 1995) under seven different prior configurations. A $BF_{10} \geq 3$ would signal support for the polarity hypothesis. Likewise, a $BF_{10} \leq \frac{1}{3}$ would constitute evidence *against* the polarity hypothesis, while $\frac{1}{3} < BF_{10} < 3$ would be an inconclusive null result.

2.5 Pilot Study, Power Analysis, and Sample Size

A pilot study with medium sample size ($N = 135$) and the same design as the presented study was initially conducted. The purpose of this pilot study was twofold: First, it served as a way of ensuring that the experimental materials would elicit plausible responses in a preliminary sample of naïve participants. Second, it allowed us to obtain a reliable estimate of the magnitude of the interaction of interest. Using that estimate, it was then possible to prospectively analyse statistical power for various possible sample sizes.

As illustrated by Figure 5, we computed the statistical power for different sample sizes ranging from $N = 50$ to $N = 500$, using simulations as implemented in the *simr* package (Green & MacLeod 2016) for the R programming language (R Core Team 2023). For each tested sample size, we ran 1,000 simulations. The minimum required sample size for obtaining statistical power of at least 0.8 appeared to be $N = 400$. Therefore, we chose to recruit exactly 400 new participants (post-exclusion) for the main experiment presented here.

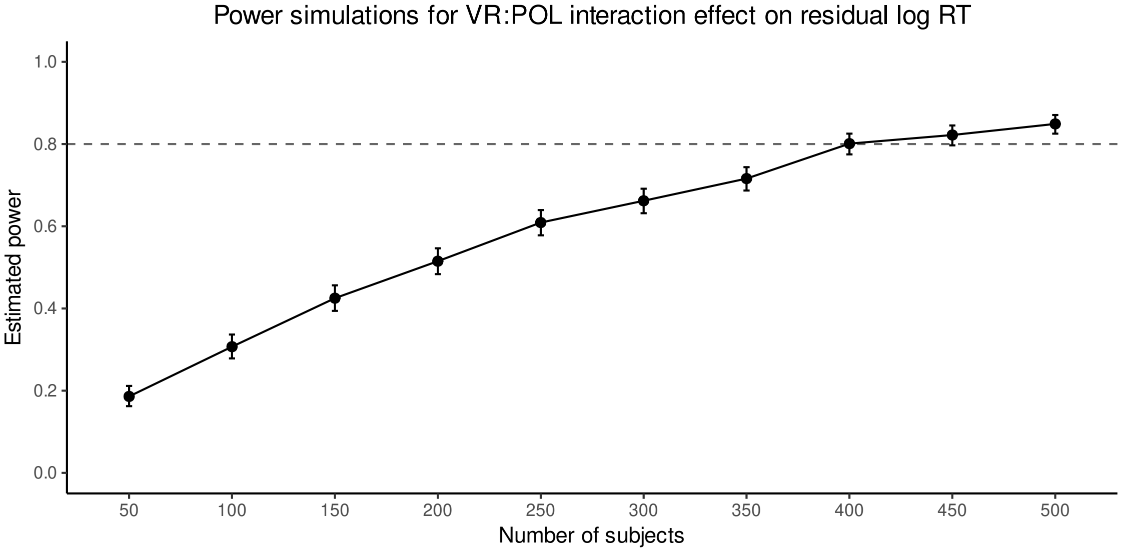


Figure 5: Prospective power for detecting `VR:POL` interaction effect (i.e., between verification response and polarity) on residual log RT, computed in order to determine a planned sample size for our main experimental study, probing different possible sample sizes from 50 to 500. The assumed effect-size estimate is drawn from pilot response data from overall 135 subjects. 1,000 simulations were run for each probed sample size. Error bars represent 95 % CIs. Apparently, meeting the desired power threshold of 0.8, visualised as a grey dashed line, requires collecting data from at least 400 subjects.

3 Results

In the collected response data from 400 participants, the predicted interaction between verification response and polarity shows no significant effect ($\hat{\beta}_{\texttt{VR:POL}} = -0.06$, $SE = 0.03$, 95 % *CrI* [−0.23, 0.11], $P(\beta_{\texttt{VR:POL}} > 0) = .2293$). This result does not provide support for the polarity hypothesis. Bayes factors provide a similar picture (see Table 5): For all of the assessed priors on the parameter of interest $\beta_{\texttt{VR:POL}}$, the BF_{10} values did not reach 3. The two most agnostic priors, $\beta_{\texttt{VR:POL}} \sim \text{Normal}(0, 1)$ and $\beta_{\texttt{VR:POL}} \sim \text{Normal}(0, 0.5)$, show BF_{10} values of 0.13 and 0.26, respectively, thus even signalling evidence in favour of the null (as $BF_{10} < \frac{1}{3}$).

Table 4 summarises all fixed-effect estimates. The only parameter with a meaningfully non-zero estimate is the intercept of residual log RT ($\hat{\alpha} = 0.35$, $SE = 0.08$, 95 % *CrI*

	$\hat{\alpha}$	*SE*	95 % *CrI*	$P(\alpha < 0)$
Intercept	0.35	0.08	[0.20, 0.51]	.0001
	$\hat{\beta}$	*SE*	95 % *CrI*	$P(\beta > 0)$
Verification response (`VR`)	−0.03	0.05	[−0.13, 0.07]	.2559
Polarity (`POL`)	−0.06	0.05	[−0.17, 0.05]	.1240
Interaction of `VR:POL`	−0.06	0.09	[−0.23, 0.11]	.2293

Table 4: Fixed-effect estimates from statistical analysis of residual log RTs, based on SPV response data collected from 400 Prolific participants.

Null	Alternative	BF_{10}
$\beta_{\texttt{VR:POL}} = 0$	$\beta_{\texttt{VR:POL}} \sim \text{Normal}(0, 1)$	0.13
$\beta_{\texttt{VR:POL}} = 0$	$\beta_{\texttt{VR:POL}} \sim \text{Normal}(0, 0.5)$	0.26
$\beta_{\texttt{VR:POL}} = 0$	$\beta_{\texttt{VR:POL}} \sim \text{Normal}(0, 0.2)$	0.57
$\beta_{\texttt{VR:POL}} = 0$	$\beta_{\texttt{VR:POL}} \sim \text{Normal}(0, 0.1)$	0.82
$\beta_{\texttt{VR:POL}} = 0$	$\beta_{\texttt{VR:POL}} \sim \text{Normal}(0, 0.05)$	0.96
$\beta_{\texttt{VR:POL}} = 0$	$\beta_{\texttt{VR:POL}} \sim \text{Normal}(0, 0.02)$	0.98
$\beta_{\texttt{VR:POL}} = 0$	$\beta_{\texttt{VR:POL}} \sim \text{Normal}(0, 0.01)$	1.00

Table 5: Results of Bayes factor analysis. Bayes factors BF_{10} were computed using the Savage–Dickey method (Verdinelli & Wasserman 1995). The same null model, which constrains the interaction of interest to zero, is compared against alternative models with differently informative priors.

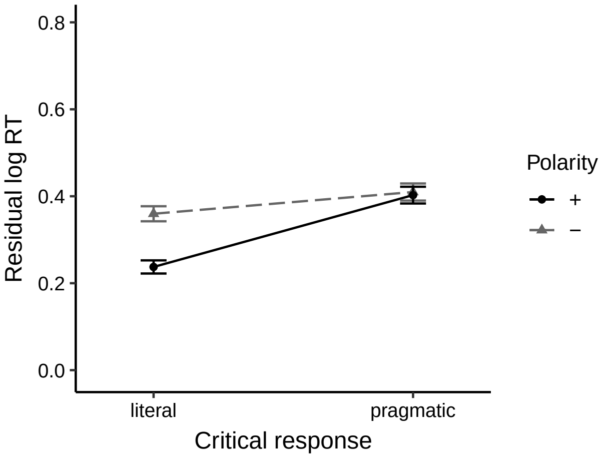


Figure 6: Mean residual log RTs grouped by (critical-trial) verification response and by polarity. The displayed error bars represent within-subject standard errors (computed following Morey et al. 2008). The underlying SPV response data was collected from 400 Prolific participants.

[0.20, 0.50], $P(\alpha < 0) = .0001$). Since the predictor variables have been sum-coded, the intercept can be interpreted as the average residual log RT, that is, as the average additional delay associated with critical trials when compared to control trials. Figure 6 visualises mean residual log RTs grouped by the levels of the predictor variables response and polarity. In Figure 7, mean residual log RTs are plotted again, but now additionally grouped by item. We observe that only the Possibility item displays a very obvious, 'X'-shaped effect pattern as predicted by the polarity hypothesis. The other three items either show only a very minor difference in slope steepness between positive and negative polarity (Quantity item), basically no such interaction (Space item), or actually a pattern where there is a slight interaction of the opposite-than-expected sign (Time item). These remarkable between-item differences suggest that time patterns of SI processing are highly sensitive to the particular scale or linguistic material involved, even after controlling for the predictors' polarity and response. Future research may probe these differences in a systematic and confirmatory way.

Figure 8 shows the ratios of 'Good' responses, grouped by scale and by whether a trial was critical (i.e., pragmatically ambiguous) or a control trial (i.e., unambiguous). As expected, the ‘Good’ ratio for the critical condition consistently falls between the ratio for the ‘Bad’-control and ‘Good’-control conditions. However, the precise ratio for the critical condition varies heavily between scales, ranging from just 31.8 % (scale +time, i.e., ⟨sometimes, always⟩) up to close-to-ceiling 86.4 % (scale +space, i.e., ⟨somewhere, everywhere⟩). Again, this points to potentially interesting between-scale differences in SI processing that are not yet accounted for.

4 Discussion

According to the scalar polarity hypothesis proposed by van Tiel et al. (2019) and further developed by van Tiel & Pankratz (2021), deriving scalar implicatures from underinformative sentences is more effortful than opting for a literal interpretation only if the underlying scale is positively polar (e.g., ⟨some, all⟩). Negatively polar scales, on the other hand, are expected to give rise to scalar implicatures that are *less* effortful (or at least equally as effortful) to derive, compared to constructing a literal interpretation. One way in which these differences in processing effort are assumed to be reflected in experimental data is through response-time patterns in verification-based paradigms, with higher effort being linked to slower responses. Therefore, the polarity hypothesis predicts a B&N effect (as found by Bott & Noveck 2004) for positively polar scales like ⟨some, all⟩, but a reversed (or at least absent) B&N effect for negatively polar scales such as ⟨not all, none⟩.

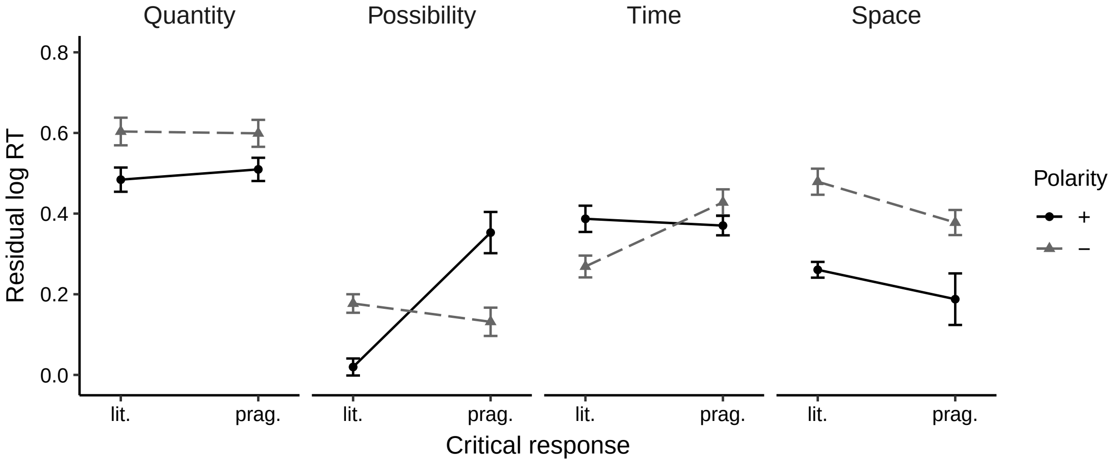


Figure 7: For each item separately: Mean residual log RTs grouped by (critical-trial) verification response and by polarity. The displayed error bars represent within-subject standard errors (computed following Morey et al. 2008). The underlying SPV response data was collected from 400 Prolific participants.

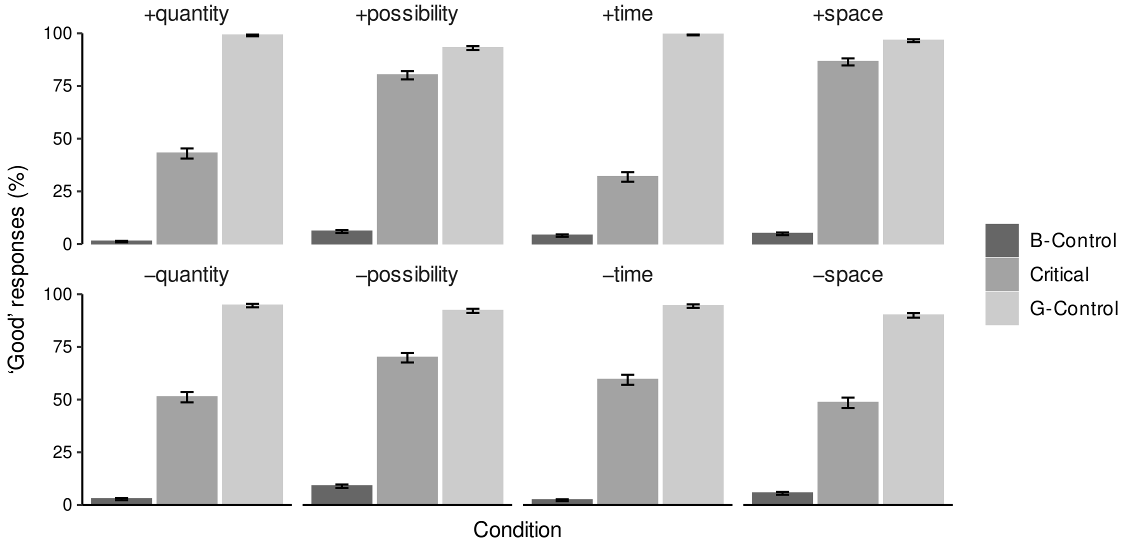


Figure 8: Bar plots showing ratios of 'Good' responses in the SPV data collected from 400 Prolific participants (post-exclusion), grouped by scale (+quantity, +possibility, ..., −space) and by three different condition set-ups: control trials where the expected response is 'Bad' (B-Control), critical trials (Critical), and control trials with a 'Good' expected response (G-Control). The displayed error bars represent within-subject standard errors (computed following Morey et al. 2008).

In an attempt to test this prediction, we designed an experiment based on sentence–picture verification that would assess SI processing across four symmetrical, positive–negative pairs of scales. Conducting this experiment in an initial pilot study with 135 participants allowed us to derive an estimate of the magnitude of the effect of interest, the interaction between response and polarity, that was then fed into a simulation-based power analysis. This power analysis indicated that, given our experimental design, at least 400 participants would be required in order to detect the desired effect with 80 % power. Hence, we preregistered our main experimental study with a sample size of $N = 400$ (post-exclusion) and conducted it accordingly.

The data collected from this main experimental study did not show the predicted interaction between response and polarity on verification times. Thus, it does not provide support for the polarity hypothesis. Nevertheless, some concerns about the generalisability of this result remain: After inspecting how the effect pattern of interest manifests itself separately in each of our four examined items (i.e., pairs of polarity-contrastive scales), we find that for what we have labelled the Possibility item (i.e., the scale pair ⟨might be, is definitely⟩—⟨might not be, is definitely not⟩), the expected interaction pattern is saliently present. The remaining three items, by contrast, display neutral or even slightly opposite effect patterns.

The present results call for future work to examine SI processing under varying polarity more broadly, perhaps with a more diverse set of scales and items, in order to ensure generalisable conclusions. It might be true that the scalar polarity hypothesis as proposed by van Tiel et al. (2019) and van Tiel & Pankratz (2021) does, indeed, apply to some types of scales and linguistic contexts. But perhaps the effect of scalar polarity is rather itself modulated by the phenomenon of scalar diversity (van Tiel et al. 2016), that is, the observation that response patterns to SIs are heavily dependent on the particular scales involved, instead of behaving in a clear-cut manner across any kind of scale.

4.1 Alternative Explanations for Variation

It has been pointed out (Doran et al. 2009; van Tiel et al. 2016) that there is a mismatch between scalar implicature in its traditional theoretical (semantic/pragmatic) conception which tacitly assumes all of its instances to display the same uniform properties, on one hand, and the large variability between scales usually detected when experimentally assessing the response behaviour of real human comprehenders, on the other hand. Van Tiel et al. argue that this *uniformity assumption* has long remained unchallenged because of a disproportionate focus on examining the scale ⟨some, all⟩, while simply presupposing that whatever holds for it should also hold for any other scale. Consequently, a new major task that emerges for researchers is to account for the apparent between-scale variability, i.e., scalar diversity, by categorising scales along various newly proposed properties. For example, van Tiel et al. come up with two broader scale properties that they call *availability* and *distinctness*: In a particular two-element scale $\langle e_1, e_2 \rangle$, availability quantifies the degree to which the stronger scalar e_2 becomes mentally activated once processing an utterance with the weaker scalar e_1. Depending on various factors like grammatical class, word frequency, or distributional semantic similarity, this degree may vary considerably across scales. Linking this to SI processing, a plausible assumption would be that in cases of high availability of the stronger scalemate, SIs are more likely to be drawn. However, even though some experimental results by Doran et al. (2009) can be interpreted in support of this assumption, van Tiel et al. (2016) themselves do not find any significant effect of availability on SI rates when putting it to the test experimentally. Yet, Hu et al. (2023) report that the expectation of alternatives estimated from a neural language model, likely a proxy for availability, successfully predicts human SI rates.

Further, given a particular scale $\langle e_1, e_2, \ldots, e_n \rangle$, distinctness describes the degree to which a stronger scalar e_k is more informative than a weaker scalar e_i, with $1 \leq i < k \leq n$. Obviously, by definition, the informativity gap (a.k.a. *semantic distance*) between e_n and e_1 is larger than the gap between e_2 and e_1 (if $n > 2$), for instance. As an illustrative example of this, consider the four-element scale ⟨some, many, most, all⟩. But even across cases where only a two-element scale seems to be the maximal possible linguistic conceptualisation of a meaning dimension of interest, one can still talk about differences in distinctness: This secondary sense of distinctness is referred to by van Tiel et al. (2016) as a scale's *boundedness*. They would argue that a scale like ⟨possible, certain⟩ is bounded because its strongest term represents a logical end point on the spanned meaning dimension, i.e., the meaning dimension of probability—'certain' means 100 % probability, and there is no such thing as, say, 101 % probability. By contrast, a scale like ⟨warm, hot⟩ is unbounded because its strongest term does not represent a logical end point on the spanned meaning dimension, i.e., of temperature. Regardless of which sense of distinctness (semantic distance or boundedness) is considered, van Tiel et al. lay out the following assumption: When a comprehender encounters the weaker term of some scale in an utterance, they are more likely to pragmatically reject the truth of a hypothetical alternative utterance with a stronger term of the same scale if the distinctness between the weaker and the stronger term is large. And, indeed, an experimental assessment of this assumption, carried out by van Tiel et al. (2016), appears to confirm that.

It would now be useful to consider if there are differences in availability or distinctness between the scales that we ourselves have examined in the present study. This may help us understand if our results are perhaps confounded by such scale differences.

Let us look at the property of distinctness first. We would argue that all of the eight scales we examined are bounded since the strongest term of each does, in fact, represent a logical end point on its respectively spanned meaning dimension. Further, all of the scales display a logical norm of correctness (Dieussaert et al. 2011) as they resemble a well-defined quantification pattern of either ⟨∃, ∀⟩ or ⟨¬∀, ¬∃⟩ in terms of truth-value distributions on their respective meaning dimensions. Hence, we also do not see any difference in distinctness in the sense of within-scale semantic distance across our examined scales, considering that the informativity gap between any ∃-like scalar term and its ∀-like stronger counterpart (or between a ¬∀-like and a ¬∃-like term) is always equally large, conceptually. Therefore, it is not plausible to object to our present findings on the basis of suspecting a confound due to differential scale distinctness.

What cannot be ruled out, though, is a confound due to the scale property of availability: Clearly, the scalars in the Quantity item are expressions of different grammatical class (they are pronouns) than the scalars in the Possibility item (modal verbal expressions) or in the Time and Space items (at their core, adverbs). Note that van Tiel et al. (2016) theorise that availability is higher in case of closed grammatical classes (say, pronouns) than it is for open grammatical classes (say, adverbs). So a basic suspicion about these differences in grammatical class potentially affecting our results is certainly justified. Also from the perspective of relative word frequency, one could argue that there are noteworthy availability differences across the examined scales: For instance, it is reasonable to assume that the phrase parts 'some of' and 'all of' occur similarly frequently in English. But the phrase part 'might be' is arguably produced and encountered much more often than 'is definitely'. Therefore, one should acknowledge a substantial difference in relative word frequency between at least two of our tested scales. This, in turn, translates again to there being a difference in availability, with scales in which the weaker scalemate is relatively highly frequent yielding SIs less often than scales where the opposite is the case, according to the reasoning by van Tiel et al. (2016). Similar arguments can probably be made with regard to the third subproperty of availability alluded to by van Tiel et al., that is, distributional semantic relatedness, but we do not attempt to show this here. Either way, we can already conclude that the availability property does, indeed, show differences across the eight scales that we tested, which makes it possible that some of our observed between-scale differences in polarity-hypothesis-relevant effect patterns are somehow correlated with such differences in availability.

Aside from the above issues related to violations of the often-held uniformity assumption regarding SIs, there is also another tacit assumption about SIs that deserves to be examined critically: the *homogeneity assumption* (as questioned by Degen 2015). Rather than supposing a lack of between-scale variability (like the uniformity assumption does), the homogeneity assumption supposes that there is no *within-scale* variability. That is, any SIs derived from the same particular scale are assumed to all behave in the same way, unaffected by differences in their linguistic contexts. Even though, already intuitively, this alleged context-independence of SIs seems somewhat hard to defend, it is, in fact, a key assumption held by major traditional theoretical frameworks which rely on it for distinguishing so-called generalised conversational implicatures (GCIs, which include SIs) from particularised conversational implicatures (PCIs). In the seminal publication by Degen (2015), this homogeneity assumption was famously challenged on the grounds of incompatible empirical evidence from a corpus-based study. Further evidence for context-dependent within-scale variability of SIs—and hence against the homogeneity assumption—is provided, e.g., by Li et al. (2021) and by Ronai & Xiang (2022). Being aware of such evidence, one may rightfully ask oneself if the results yielded by our present work would have looked different if the examined scales had been embedded in other linguistic or visual contexts than the particular contexts we happened to choose here when we designed our experiment.

A different concern regards the results of our present experimental study and their comparability to previous studies like those by van Tiel et al. (2019) or van Tiel & Pankratz (2021): While our sentence–picture verification data displays an average response time of 2,021 ms, the related studies by van Tiel and colleagues display average response times of only 1,232 ms and 1,267 ms, respectively. Therefore, it might be that our failure to find support for the polarity hypothesis is related to the unusually long response times observed in our data, which may have masked the expected effect to some extent. A possible explanation for this discrepancy in average response times is that our experimental study featured explicitly negative scalar terms (e.g., 'not everywhere') which may take rather long to process compared to the implicitly negative scalar expressions (e.g., 'absent') tested by van Tiel and colleagues, even on the baseline level where there is no pragmatic ambiguity involved.

5 Conclusion

Overall, our present experimental results do not yield support for the scalar polarity hypothesis proposed by van Tiel et al. (2019) and van Tiel & Pankratz (2021). Moreover, substantial between-scale differences in effect patterns remain unaccounted for. Future work may attempt to probe these differences in a confirmatory setting. If they persist, then more fine-grained theoretical accounts of how SI processing relates to cognitive effort are clearly needed. More generally, the present findings vindicate the observation of scalar diversity raised by van Tiel et al. (2016): The degree to which processing signatures of different scales can vary still remains underexplored and underappreciated. Despite the existing plethora of studies on SI processing, an even broader range of experiments is still needed to reliably assess the validity of particular theoretical accounts of SI processing (such as the polarity hypothesis). Only that will make it possible to eventually build towards an integrated model of pragmatically enriched language comprehension.

Abbreviations

SI = scalar implicature, SPV = sentence–picture verification

Data Availability (anonymised URLs while under review!)

Experimental data files and analysis scripts can be accessed here:

- https://osf.io/6bn9z/?view_only=557995afbefd40d188e34597fd7be825

Further, an anonymised version of the reported experiment's preregistration protocol is available here:

- https://pdfhost.io/v/6uiSeVhqz_ANONYMISED_Protocol_Exp2_230405

The following OSF registry contains the time-stamped preregistration of the present study:

- https://osf.io/dhpzq?view_only=ce9fa3c53ec9431b9e9a42f1a40e1828

Ethics and Consent

An ethics-and-consent statement will be provided once the submission does not have to be anonymised anymore.

Funding Information

Funding information will be provided once the submission does not have to be anonymised anymore.

Competing Interests

The authors have no competing interests to declare.

References

Bill, Cory & Romoli, Jacopo & Schwarz, Florian. 2018. Processing presuppositions and implicatures: Similarities and differences. *Frontiers in communication* 3. 44. https://doi.org/10.3389/fcomm.2018.00044

Bott, Lewis & Noveck, Ira A. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language* 51(3). 437–457. https://doi.org/10.1016/j.jml.2004.05.006

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software* 80. 1–28. https://doi.org/10.18637/jss.v080.i01

Chierchia, Gennaro et al. 2004. Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and beyond* 3. 39–103.

Clark, Herbert H & Chase, William G. 1972. On the process of comparing sentences against pictures. *Cognitive psychology* 3(3). 472–517. https://doi.org/10.1016/0010-0285(72)90019-9

Cremers, Alexandre & Chemla, Emmanuel. 2014. Direct and indirect scalar implicatures share the same processing signature. In *Pragmatics, semantics and the case of SIs*, 201–227. Springer. https://doi.org/10.1057/9781137333285_8

De Neys, Wim & Schaeken, Walter. 2007. When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental psychology* 54(2). 128. https://doi.org/10.1027/1618-3169.54.2.128

Degen, Judith. 2015. Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and pragmatics* 8. 11. https://doi.org/10.3765/sp.8.11

Degen, Judith & Tanenhaus, Michael K. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive science* 39(4). 667–710. https://doi.org/10.1111/cogs.12171

Dieussaert, Kristien & Verkerk, Suzanne & Gillard, Ellen & Schaeken, Walter. 2011. Some effort for some: Further evidence that scalar implicatures are effortful. *Quarterly journal of experimental psychology* 64(12). 2352–2367. https://doi.org/10.1080/17470218.2011.588799

Doran, Ryan & Baker, Rachel & McNabb, Yaron & Larson, Meredith & Ward, Gregory. 2009. On the non-unified nature of scalar implicature: An empirical investigation. *International review of pragmatics* 1(2). 211–248. https://doi.org/10.1163/187730909X12538045489854

Fodor, Jerold A & Garrett, Merrill F. 1975. The psychological unreality of semantic representations. *Linguistic inquiry* 6(4). 515–531.

Green, Peter & MacLeod, Catriona J. 2016. simr: An R package for power analysis of generalised linear mixed models by simulation. *Methods in ecology and evolution* 7(4). 493–498. https://doi.org/10.1111/2041-210X.12504

Grice, Herbert P. 1975. Logic and conversation. In *Speech acts*, 41–58. Brill.

Horn, Laurence Robert. 1972. *On the semantic properties of logical operators in English*. University of California, Los Angeles.

Hu, Jennifer & Levy, Roger & Degen, Judith & Schuster, Sebastian. 2023. Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics* 11. 885–901. https://doi.org/10.1162/tacl_a_00579

Huang, Yi Ting & Spelke, Elizabeth & Snedeker, Jesse. 2013. What exactly do numbers mean? *Language learning and development* 9(2). 105–129. https://doi.org/10.1080/15475441.2012.658731

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press. https://doi.org/10.7551/mitpress/5526.001.0001

Lewandowski, Daniel & Kurowicka, Dorota & Joe, Harry. 2009. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis* 100(9). 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Li, Elissa & Schuster, Sebastian & Degen, Judith. 2021. Predicting scalar inferences from "or" to "not both" using neural sentence encoders. In *Proceedings of the Society for Computation in Linguistics 2021*. 446–450. https://doi.org/10.7275/xr01-a852

Marty, Paul & Chemla, Emmanuel & Spector, Benjamin. 2013. Interpreting numerals and scalar items under memory load. *Lingua* 133. 152–163. https://doi.org/10.1016/j.lingua.2013.03.006

Marty, Paul & Romoli, Jacopo & Sudo, Yasutada & van Tiel, Bob & Breheny, Richard. 2020. Processing implicatures: A comparison between direct and indirect SIs. Oral presentation at Experiments in Linguistic Meaning (ELM), Philadelphia, PA.

Morey, Richard D et al. 2008. Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in quantitative methods for psychology* 4(2). 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Noveck, Ira A. 2001. When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition* 78(2). 165–188. https://doi.org/10.1016/s0010-0277(00)00114-1

Pusse, Florian & Sayeed, Asad & Demberg, Vera. 2016. Lingoturk: Managing crowdsourced tasks for psycholinguistics. In *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: Demonstrations*. 57–61. https://doi.org/10.18653/v1/N16-3012

R Core Team. 2023. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Rips, Lance J. 1975. Quantification and semantic memory. *Cognitive psychology* 7(3). 307–340. https://doi.org/10.1016/0010-0285(75)90014-6

Romoli, Jacopo & Schwarz, Florian. 2015. An experimental comparison between presuppositions and indirect scalar implicatures. In *Experimental perspectives on presuppositions*, 215–240. Springer. https://doi.org/10.1007/978-3-319-07980-6_10

Ronai, Eszter & Xiang, Ming. 2022. Quantifying semantic and pragmatic effects on scalar diversity. *Proceedings of the Linguistic Society of America* 7(1). 5216. https://doi.org/10.3765/plsa.v7i1.5216

Sperber, Dan & Wilson, Deirdre. 1986. *Relevance: Communication and cognition*. Blackwell.

Tavano, Erin & Kaiser, Elsi. 2010. Processing scalar implicature: What can individual differences tell us? *University of Pennsylvania working papers in linguistics* 16(1). 24.

van Tiel, Bob & Pankratz, Elizabeth. 2021. Adjectival polarity and the processing of scalar inferences. *Glossa: A journal of general linguistics* 6(1). https://doi.org/10.5334/gjgl.

1457

van Tiel, Bob & Pankratz, Elizabeth & Sun, Chao. 2019. Scales and scalarity: Processing scalar inferences. *Journal of memory and language* 105. 93–107. https://doi.org/10.1016/j.jml.2018.12.002

van Tiel, Bob & van Miltenburg, Emiel & Zevakhina, Natalia & Geurts, Bart. 2016. Scalar diversity. *Journal of semantics* 33(1). 137–175. https://doi.org/10.1093/jos/ffu017

Verdinelli, Isabella & Wasserman, Larry. 1995. Computing bayes factors using a generalization of the savage-dickey density ratio. *Journal of the American Statistical Association* 90(430). 614–618. https://doi.org/10.1080/01621459.1995.10476554