

## Scalar inferences and cognitive load<sup>1</sup>

Bob VAN TIEL — *Leibniz-Zentrum Allgemeine Sprachwissenschaft*

Paul Marty — *Leibniz-Zentrum Allgemeine Sprachwissenschaft*

Elizabeth PANKRATZ — *Leibniz-Zentrum Allgemeine Sprachwissenschaft*

Chao SUN — *Leibniz-Zentrum Allgemeine Sprachwissenschaft*

**Abstract.** A number of studies have found that participants are less likely to interpret the scalar words ‘some’ and ‘or’ with an upper bound when their cognitive resources are burdened, thus suggesting that the computation of scalar inferences is cognitively effortful. We conducted two sentence-picture verification tasks to determine whether this finding generalises to other scalar words. In Exp. 1, we manipulated cognitive load by asking participants to memorise simple or complex grid patterns during the experiment (cf. De Neys and Schaeken, 2007). In Exp. 2, we manipulated cognitive load by varying the time participants could take to process the sentences and pictures (cf. Chevallier et al., 2008). In this way, we tested seven scalar words: ‘some’, ‘or’, ‘low’, ‘scarce’, ‘might’, ‘most’, and ‘try’. We expected to find lower rates of scalar inferences when participants experienced greater cognitive load, i.e., when they had to memorise complex grid patterns in Exp. 1, and when they had less processing time available in Exp. 2. We find significant effects of memory load in the expected direction, but only for positively scalar words, i.e., for scalar words that denote a lower bound. We fail to find any significant effects of processing time. We explain these findings by arguing that the scalar inferences of positively scalar words introduce negative information into the meaning of the sentence, and that the processing of such negative information is cognitively demanding.

**Keywords:** scalar inferences, experimental pragmatics, working memory.

### 1. Introduction

It is generally assumed that the *literal* meaning of scalar words, such as ‘some’ and ‘or’, is lower-bounded only. Thus, the literal meaning of (1) can be paraphrased as ‘I ate at least some and possibly all of the pie’.

(1) I ate some of the pie.

At the same time, it is clear that someone who utters (1) may imply that she did not eat all of the pie. This *scalar inference* is often explained as a conversational implicature along the following lines: someone who utters (1) could have been more informative by saying ‘I ate all of the pie’. Why didn’t she? Presumably because she did not eat all of the pie. In this way, ‘some’ acquires its *pragmatic* meaning as ‘at least some but not all’ (e.g., Horn, 1972; Gazdar, 1979).

At the theoretical level, then, the derivation of scalar inferences involves a protracted reasoning process that uses the literal interpretation in its premises. There has been a substantial amount of debate as to whether this reasoning process is reflected in the cognitive processing of scalar inferences, i.e., whether the literal meaning of scalar words is easier to retrieve than the pragmatic meaning.

---

<sup>1</sup>We would like to thank Paul Marty for his suggestion to test different types of grid patterns in the memory load task. In addition, we would like the organisers, reviewers, and participants of *Sinn und Bedeutung* 23, in particular Marieke Schouwstra and Fausto Carcassi, for their valuable feedback.

On the one hand, proponents of *relevance theory* have argued that hearers initially interpret utterances with scalar words literally. If the hearer is not satisfied with the relevance of this literal interpretation, she may choose to make it more relevant by computing the scalar inference. This process of pragmatic enrichment is assumed to be cognitively effortful (e.g., Sperber and Wilson, 1986). On the other hand, Levinson (2000) made a case for the primacy of the pragmatic interpretation. Levinson argues that scalar words are standardly interpreted with an upper bound. In certain circumstances, this upper bound may be cancelled to arrive at the literal interpretation. According to Levinson, this process of cancellation is cognitively effortful.

In other words, whereas relevance theory argues that the computation of scalar inferences is associated with a processing cost, Levinson's defaultist approach holds that it is rather the cancellation of scalar inferences that is cognitively effortful.

Relevance theory and Levinson's defaultist approach make a number of conflicting empirical predictions. One such prediction centers on the effect of cognitive load on the probability of deriving scalar inferences. If the computation of scalar inferences draws upon cognitive resources, as relevance theory holds, people should be *less* likely to compute scalar inferences when their cognitive resources are burdened; by contrast, if it is the cancellation of scalar inferences that is cognitively effortful, people should be *more* likely to compute scalar inferences.

### 1.1. Prior work

To test these predictions, De Neys and Schaeken (2007) conducted a sentence verification task in which participants had to provide truth judgements to underinformative sentences containing the scalar word 'some', such as (2) (cf. also Bott and Noveck, 2004).

- (2)    a.    Some dogs are mammals.  
       b.    Some parrots are birds.

These sentences are true on their literal interpretation but false if the scalar inference is computed and 'some' is interpreted as excluding 'all'. Hence, 'true' responses indicate that participants arrived at a literal interpretation; 'false' responses suggest that a scalar inference was computed. In what follows, we conveniently refer to these two types of responses as *literal* and *pragmatic*, respectively.

While providing their truth judgements, participants in De Neys and Schaeken's experiment had to memorise dot patterns in  $3 \times 3$  matrices. These dot patterns were either *simple*, consisting of three dots in a horizontal or vertical line, or *complex*, consisting of four dots scattered across the matrix. In this way, De Neys and Schaeken manipulated the degree of cognitive load that participants experienced while they evaluated the target sentences.

In line with the relevance-theoretic predictions, participants were less likely to respond pragmatically if they had to memorise complex dot patterns (73%) compared to simple ones (79%). Thus, greater cognitive load decreased the probability that participants computed scalar inferences. De Neys and Schaeken's results have since been replicated in at least three studies (Dieussaert et al., 2011; Marty and Chemla, 2013; Marty et al., 2013).

Chevallier et al. (2008) provide a second piece of evidence in favour of relevance theory. Their study focuses on the interpretation of 'or' rather than 'some'. Participants in their Exp. 1 were

presented with strings of letters that were described by means of sentences such as (3).

(3) There is an A or a B.

Participants had to indicate if the sentence was a correct description of the letter string. In the target condition, the corresponding letter string contained both an A and a B, thus verifying the sentence on its literal interpretation but falsifying the scalar inference.

There were three versions of the experiment. In the *fast* version, participants saw the letter string for one second. Afterwards, the letter string disappeared and was replaced by the sentence. The *normal* version was identical to the fast version, except that the letter string remained on screen when the sentence was presented. The *slow* version was identical to the normal version, except that participants had to wait for three seconds after the presentation of the sentence before they were allowed to respond. In this way, Chevallier et al. manipulated the time and effort that participants could invest in the verification task.

In line with the relevance-theoretic predictions, participants were more likely to respond pragmatically when they had more time to process the sentences. Thus, in the fast version, participants responded pragmatically 20% of the time; in the normal version, 25%, and in the slow version, 48%.

## 1.2. Broadening the scope

The memory load effect of De Neys and Schaeken and the processing time effect of Chevallier et al. have been shown for two scalar words only: ‘some’ and ‘or’. However, if these effects are to be taken as evidence for relevance theory, rather than simply as refuting Levinson’s defaultist approach, they should generalise across the entire family of scalar words. In recent work, we have shown that this *uniformity assumption* may not hold, at least when it concerns De Neys and Schaeken’s memory load effect (van Tiel et al., 2019).

Van Tiel et al. tested seven scalar words: ‘low’, ‘scarce’, ‘might’, ‘or’, ‘some’, ‘most’, and ‘try’. For each of the seven scalar words, van Tiel et al. constructed a sentence and three types of pictures: one in which the sentence was unambiguously true, one in which it was unambiguously false, and one in which the sentence was literally true but false if the corresponding scalar inference was derived. The first two picture types constitute the control condition; the third picture type the target condition. Table 1 shows example sentences and pictures for each scalar word.

Participants were presented with sentences and pictures, and they had to indicate if the sentence adequately described the picture. In Exp. 2, van Tiel et al. manipulated the degree of cognitive load that participants experienced in a similar way to De Neys and Schaeken (2007). Thus, participants were assigned to one of three conditions: in the no-load condition, participants did not experience any cognitive load; in the low-load condition, participants had to memorise simple patterns consisting of three horizontally aligned black squares in a 3×3 matrix; in the high-load condition, participants had to memorise more complex patterns consisting of four black squares in a 3×3 matrix. See Fig. 1 for example grids.

Van Tiel et al. observed significant negative effects of cognitive load for ‘might’, ‘or’, ‘some’, ‘most’ on the probability that participants responded pragmatically in the target condition. Participants were thus less likely to derive the scalar inferences associated with these scalar words






















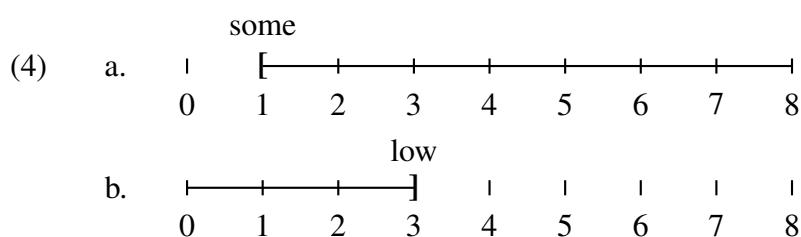
<i>Sentence</i>	<i>Control (T)</i>	<i>Control (F)</i>	<i>Target</i>
The battery is low.			
Red flowers are scarce.			
Either the apple or the pepper is red.			
The arrow might land on red.			
Some of the socks are pink.			
Most of the apples are green.			
He tried to tie his tie.			

Table 1: Sentences and example pictures for each scalar term from van Tiel *et al.* (2019).

when they were under greater cognitive load. However, the probability of pragmatic responses for ‘low’, ‘scarce’, and ‘try’ was independent of the degree of cognitive load.

Van Tiel *et al.* explain these findings based on the notion of *scalarity*. Scalar words are either *positively* or *negatively* scalar depending on whether they denote a lower or upper bound on their dimension (e.g., Horn, 1989; Matsumoto, 1995). Thus, ‘some’ is positively scalar because it denotes a lower bound: if the pie from example (1) consists of eight slices, the meaning of ‘some of the pie’ can be visualised as in (4a). Conversely, ‘low’ is negatively scalar because it denotes an upper bound on its dimension. Hence, the meaning of ‘low on pie’ can be visualised as in (4b).



The notion of scalarity is akin to that of *monotonicity*. Hence, another way of bringing out the

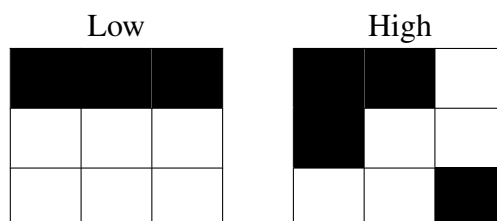


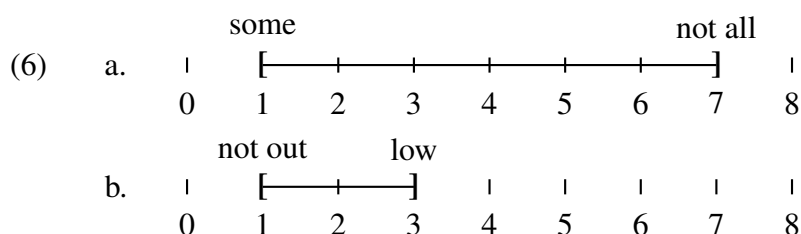
Figure 1: Examples of low-load and high-load matrices that participants had to memorise in van Tiel et al. (2019: Exp. 2).

contrast between ‘some’ and ‘low’ is by inspecting their inferential potential: ‘some’ allows for inferences from a set to its superset, as shown by the validity of the argument in (5a); ‘low’ allows for inferences from a set to its subset, as shown by the validity of the argument in (5b). In both cases, the argument becomes invalid if the premise and conclusion are reversed.

- (5) a. I ate some of the apple pie.  $\Rightarrow$  I ate some of the pie.
- b. We are low on pie.  $\Rightarrow$  We are low on apple pie.

‘Scarce’ patterns with ‘low’ in that it is negatively scalar; ‘or’, ‘might’, ‘most’, and ‘try’ are positively scalar, just like ‘some’.

The scalarity of a word determines, among other things, the polarity of the corresponding scalar inference: positively scalar words give rise to negative, i.e., upper-bounding scalar inferences; negatively scalar words to positive, i.e., lower-bounding scalar inferences. Thus, the pragmatically enriched meanings of ‘some of the pie’ implying ‘not all of the pie’ and ‘low on pie’ implying ‘not out of pie’ can be visualised as in (6).



Note that the negativity of the scalar inferences of positively scalar words is not reflected in any negative elements in the surface form of the scalars; rather, their negativity is implicit, involving the placement of an upper bound on the dimension over which the scalar word quantifies (cf. Horn, 1989: p. 188ff.).

Van Tiel et al. argue that the scalar inferences of positively scalar words are cognitively demanding because they introduce a negative proposition into the meaning of the sentence. There is a substantial body of evidence showing that the processing of negative information is cognitively effortful (e.g., Clark and Chase, 1973; Deschamps et al., 2015; Geurts et al., 2010). Hence, the derivation of the scalar inferences of positively scalar words—but not negatively scalar ones—is associated with a processing cost.

There are, however, three observations that sit uneasily with this explanation. First, ‘try’, which is positively scalar, did not show a significant effect of cognitive load. Second, the cognitive load effects for the other positively scalar words were found primarily in the comparison be-

tween the no-load and low-load conditions, rather than between the low-load and high-load conditions, which is what De Neys and Schaeken found for ‘some’. Third, the effect of memory load on the probability of deriving the scalar inferences of positively scalar words did not always differ significantly from its effect on negatively scalar words.

### 1.3. Current study

One possible explanation for this less than perfectly consistent pattern of results is that the difference in complexity between the grids in the low-load and high-load conditions was not sufficiently pronounced to affect the probability of pragmatic responses. Work in progress with Paul Marty suggests that the effect of memory load on the probability of pragmatic responses becomes more robust if the difference in complexity between the low-load and high-load conditions becomes more pronounced. In order to evaluate this explanation, and to obtain a better understanding of the effect of cognitive load on scalar inferences, we conducted two sentence-picture verification tasks.

Exp. 1 replicated van Tiel *et al.*’s Exp. 2 with grids that differed more prominently in complexity. Specifically, participants were assigned to one of two conditions: in the *minimal-load* condition, participants had to memorise patterns consisting of one black square in a  $2 \times 2$  matrix; in the *maximal-load* condition, participants had to memorise patterns consisting of four black squares in a  $4 \times 4$  matrix. Fig. 2 shows example grids from both conditions.

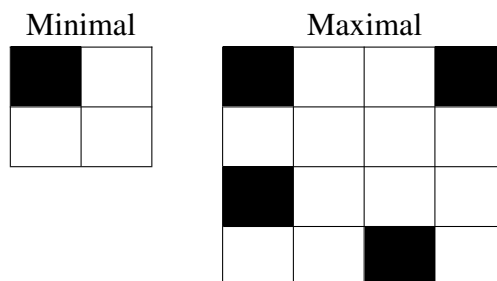


Figure 2: Examples of minimal-load and maximal-load matrices that participants had to memorise in Exp. 1.

Afterwards, as a second test of the effect of cognitive load on the derivation of scalar inferences, Exp. 2 investigates the generalisability of the results found by Chevallier *et al.* by testing the same seven scalar words as in Exp. 1 within their experimental paradigm. That is, three versions of the sentence-picture verification task were constructed: in the fast version, the picture was briefly presented and then replaced by the sentence; in the normal version, the picture remained on screen when the sentence was presented, and in the slow version, participants had to wait for three seconds before registering their truth judgements.

Taken together, these two experiments provide an insight into the effect of cognitive load on the derivation of scalar inferences. We distinguish two possible hypotheses. On the naive view that the processing of all varieties of scalar inferences should proceed along the same lines, it is expected that increased cognitive load should decrease the probability of deriving the scalar inferences of all seven scalar words. If, however, van Tiel *et al.*’s scalarity-based explanation is on the right track, it is expected that increased cognitive load only decreases the probability of

deriving the scalar inferences of the positively scalar words ‘might’, ‘or’, ‘some’, ‘most’, and ‘try’, but not the negatively scalar words ‘low’ and ‘scarce’.

In the next sections, we describe the two experiments in more detail.

## 2. Experiment 1

### 2.1. Participants

100 participants (mean age: 36, standard deviation: 10, range: 21–71, 38 females) were drafted on Mechanical Turk and were paid \$2.00 for their participation. Participants were asked to indicate their native language, but payment was not contingent on their response to this question. All of the participants indicated that they were native speakers of English.

### 2.2. Materials

The materials were the same as in van Tiel et al. (2019: Exp. 2).

That is, the experiment tested seven scales: ⟨low, empty⟩, ⟨scarce, absent⟩, ⟨or, and⟩, ⟨may, must⟩, ⟨might, will⟩, ⟨some, all⟩, ⟨most, all⟩, and ⟨try, succeed⟩. Each scale was associated with one sentence with the weaker scalar term. These sentences were paired with three types of pictures. In one picture type, the sentence was unambiguously true (‘true’ control condition), in one picture type, it was unambiguously false (‘false’ control condition), and in one picture type, the truth value of the sentence depended on whether the corresponding scalar inference was computed (target condition). That is, the sentence was true if it was interpreted literally but false if the corresponding scalar inference was computed. There were three slightly different tokens of each type of picture. Table 1 shows the seven sentences and example tokens of each picture type. The order of the items was completely randomised for each participant.

Each trial started with the presentation of a pattern of black squares in a matrix. The patterns in the minimal-load condition consisted of one black square in a  $2 \times 2$  matrix. The patterns in the maximal-load condition consisted of four black squares in a  $4 \times 4$  matrix. The black squares were never horizontally or vertically contiguous. Fig. 2 shows example matrices from both conditions.

### 2.3. Procedure

Each trial started with the presentation of a matrix, which appeared on screen for 1,200 msec. Participants were instructed to memorise the pattern in these matrices. Afterwards, a sentence and a picture were presented in the middle of the screen. Participants had to decide whether or not the sentence was a good description of the depicted situation. They could register their decision by pressing either ‘1’ (good description) or ‘2’ (bad description) on their keyboard. Once they had registered their decision, they saw an empty matrix and had to recreate the pattern that was presented at the start of the trial. To this end, participants could fill or unfill squares in the matrix by clicking on them. No feedback was given on their performance.

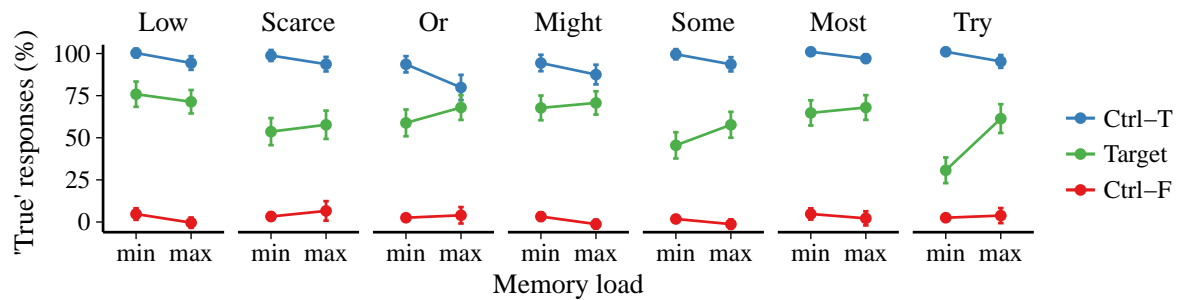


Figure 3: Percentage of ‘true’ responses for each scalar term, condition, and memory load (min = minimal load, max = maximal load).

#### 2.4. Data treatment

11 participants were removed for making mistakes in more than 20% of the control items. The mean error rate on control items of the remaining participants was 3.4% in the maximal-load condition and 1.6% in the minimal-load condition. In addition, we removed three participants from the maximal-load condition because they correctly recalled fewer than 10% of the matrices. The mean error rate on the matrix recall task was 45% in the maximal-load condition and 2.8% in the minimal-load condition. 86 participants were thus included in the analyses.

We removed items with a response time below 500 milliseconds or above 15 seconds, assuming that these correspond to accidental button presses or a lack of concentration on the task at hand (0.8% of the data).

#### 2.5. Memory load

Fig. 3 shows the percentages of ‘true’ responses for each scalar term, condition, and memory load. In the control condition, performance was close to ceiling (all error rates < 11%). The only apparent exception was the ‘true’ control condition of ‘or’, for which performance dropped from 4.3% errors in the minimal-load condition to 13.2% in the maximal-load condition (cf. Chevallier *et al.*, 2010, for similar findings).

In order to determine whether there were significant effects of memory load on the probability of pragmatic answers in the target condition, we constructed, for each scalar word, a mixed effects logistic regression model predicting responses in the target condition (literal or pragmatic) based on memory load, including random intercepts for participants and items, which was the maximal converging model for most of the scalar words. There were significant negative effects of memory load for ‘or’ ( $\beta = 2.55$ ,  $SE = 1.15$ ,  $t = 2.21$ ,  $p = .027$ ), ‘some’, ( $\beta = 2.21$ ,  $SE = 1.13$ ,  $t = 1.97$ ,  $p = .049$ ), and ‘try’ ( $\beta = 3.13$ ,  $SE = 1.18$ ,  $t = 2.65$ ,  $p = .008$ ), but not for any of the remaining scalar words (all  $t$ 's < 1.3).

To obtain a more complete picture, we also conducted an analysis in which we included the data reported by van Tiel *et al.* (2019: Exp. 2).<sup>2</sup> Thus, we have a data set with data from 250 participants who did the same sentence-picture verification task under five levels of cognitive load:

<sup>2</sup>These data can be found at: <https://data.mendeley.com/datasets/zpfm55nr33/1>



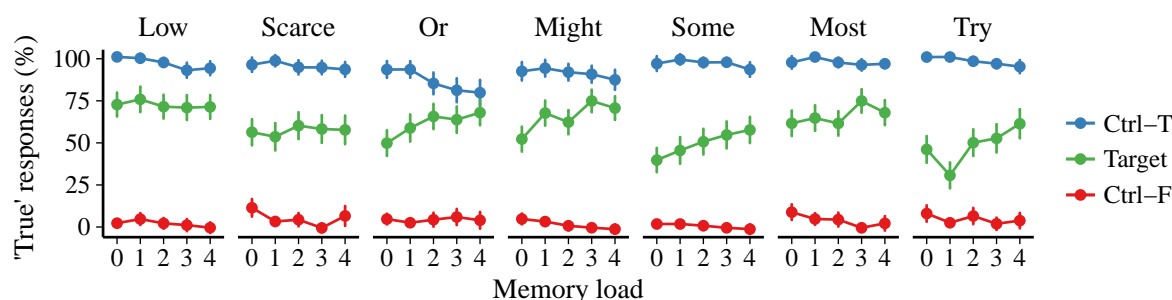


Figure 4: Percentage of ‘true’ responses for each scalar term, condition, and memory load. 0 = no load, 1 = minimal load, 2 = low load, 3 = high load, 4 = maximal load. The results from the no-load, low-load, and high-load conditions are taken from van Tiel et al. (2019: Exp. 2). Error bars represent standard errors of the mean.

no load, minimal load, low load, high load, and maximal load. Fig. 4 shows the percentages of ‘true’ responses for each scalar term, condition, and memory load.

Again, in order to determine whether there were significant effects of memory load on the probability of pragmatic answers in the target condition, we constructed, for each scalar word, a mixed effects logistic regression model predicting responses in the target condition (literal or pragmatic) based on memory load (no, minimal, low, high, or maximal), including random intercepts for participants and items, which was the maximal converging model for most of the scalar words. There were significant negative linear effects of memory load for ‘or’ ( $\beta = 2.99$ ,  $SE = 1.00$ ,  $t = 2.99$ ,  $p = .003$ ), ‘might’, ( $\beta = 10.97$ ,  $SE = 1.47$ ,  $t = 7.44$ ,  $p < .001$ ), ‘some’ ( $\beta = 2.64$ ,  $SE = 0.92$ ,  $t = 2.88$ ,  $p = .004$ ), ‘most’ ( $\beta = 1.80$ ,  $SE = 0.91$ ,  $t = 1.98$ ,  $p = .048$ ), and ‘try’ ( $\beta = 2.68$ ,  $SE = 0.91$ ,  $t = 2.94$ ,  $p = .003$ ), but not the other two scalar words (both  $t$ 's  $< 1$ ).

To determine if the effect of memory load differed across scalar words, we constructed, for the target condition of each pair of scalar words, a generalised mixed effects logistic regression model predicting response (‘true’ or ‘false’) on the basis of memory load (no, minimal, low, high, or maximal), scalar word, and their interaction. Again, these analyses only included random intercepts for participants due to convergence issues. The significance of the interactions between memory load and scalar term is provided in Table 2.

The results of this analysis largely confirm the results of the previous analyses: there was a significantly stronger negative effect of memory load on the probability of pragmatic responses for ‘or’, ‘might’, ‘some’, and ‘try’ than for ‘low’ and ‘scarce’; none of the other comparisons were statistically significant.

In summary, the results of Exp. 1 indicate that, for the positively scalar words ‘or’, ‘some’, and ‘try’, participants were significantly less likely to respond pragmatically when they had to memorise complex matrices than simple ones; no such effect was found for the negatively scalar words ‘low’ and ‘scarce’. Perhaps surprisingly, though, the positively scalar words ‘might’ and ‘most’ patterned with ‘low’ and ‘scarce’. However, once we also included into our analyses the data from van Tiel et al. (2019: Exp. 2), a more consistent pattern emerged, with all of the positively scalar words showing a significant effect of memory load in the expected direction, and no effect of memory load on their negatively scalar counterparts. Indeed, the effect of

	Scarce	Or	Might	Some	Most	Try
Low	< 1	2.97 **	3.23 **	2.65 **	1.77 ·	3.87 ***
Scarce		2.40 *	2.71 **	2.56 *	1.20	2.99 **
Or			< 1	< 1	1.28	< 1
Might				< 1	1.85 ·	< 1
Some					1.38	< 1
Most						1.37

Table 2:  $Z$  and  $p$  values indicating whether the interaction between scalar term and memory load had a significant effect on responses in the target condition for each pair of scalar terms. Note: · indicates significance at the .10 level; \* at the .05 level; \*\* at the .01 level; \*\*\* at the .001 level.

memory load on the response patterns for the positively scalar words ‘or’, ‘might’, ‘some’, and ‘try’—but not ‘most’—differed significantly from its effect on the response patterns for the negatively scalar words ‘low’ and ‘scarce’.

### 3. Experiment 2

#### 3.1. Participants

150 participants (mean age: 34, standard deviation: 10, range: 17–69, 60 females) were drafted on Mechanical Turk and were paid \$2.00 for their participation. Participants were asked to indicate their native language, but payment was not contingent on their response to this question. Two participants were removed from the analyses for having a native language other than English.

#### 3.2. Materials

The materials were the same as for Exp. 1. However, participants in Exp. 2 did not have to memorise any grid patterns during the sentence-picture verification task.

#### 3.3. Procedure

The procedure was analogous to Chevallier *et al.* (2010: Exp. 1).

Participants were presented with sentences and pictures, and they had to decide whether or not the sentence was a good description of the depicted situation. They could register their decision by pressing either ‘1’ (good description) or ‘0’ (bad description) on their keyboard.

There were three versions of the experiment. In the *fast* version, trials started with the presentation of the picture in the middle of the screen. After one second, the picture disappeared and was replaced by the sentence. The sentence remained on screen until participants registered their truth judgements. The *normal* version was identical to the fast version, except that the picture remained on screen when the sentence was presented. The *slow* version was identical to the normal version, except that participants had to wait for three seconds after sentence onset before providing their truth judgements. If they pressed one of the response buttons before three

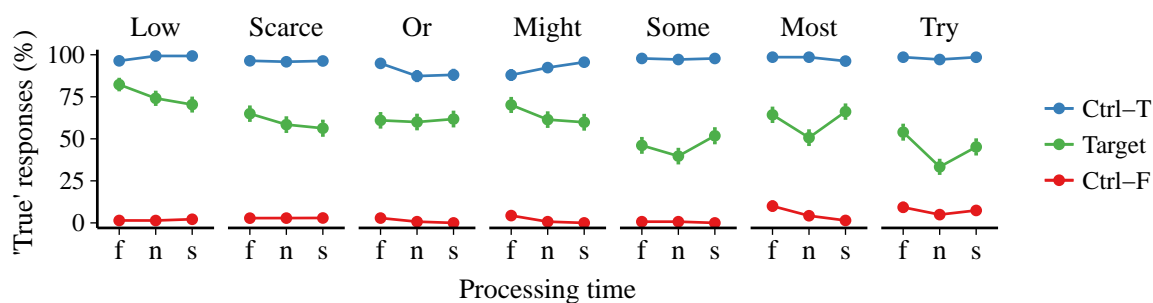


Figure 5: Percentage of 'true' responses for each scalar term, condition, and version. f = fast, n = normal, s = slow. Error bars represent standard errors of the mean.

seconds had passed, the message 'Too fast!' appeared on screen and remained there for three seconds.

### 3.4. Data treatment

8 participants were removed for making mistakes in more than 20% of the control items. The mean error rate on control items of the remaining participants was 3.6%. 140 participants were thus included in the analyses.

We removed items with a response time below 200 milliseconds or above 10 seconds, assuming that these correspond to accidental button presses or a lack of concentration on the task at hand (2.1% of the data).

### 3.5. Processing time

Fig. 4 shows the percentages of 'true' responses in the target condition for each scalar term and version (fast, normal, or slow). In order to determine if there were significant effects of version on the probability of pragmatic answers in the target condition, we constructed, for each scalar word, a mixed effects logistic regression model predicting responses in the target condition (literal or pragmatic) on the basis of version (fast, normal, or slow), including random intercepts for participants and items, which was the maximal converging model for most of the scalar words. There were no significant linear effects of version for any of the scalar words: 'might' ( $\beta = -1.49$ ,  $SE = 0.98$ ,  $Z = -1.51$ ,  $p = .130$ ), 'try', ( $\beta = -0.82$ ,  $SE = 0.75$ ,  $Z = -1.09$ ,  $p = .275$ ), nor any of the remaining scalar words (all  $Z$ 's of the remaining words being  $< 1$ ).

The results of Exp. 2 thus contradict previous findings by Chevallier et al. (2008), who found that participants were significantly more likely to interpret 'or' with an upper bound in the slow condition compared to both the fast and normal conditions. We did not find reliable effects of processing time for any of the scalar words that we tested, including 'or'.

## 4. General discussion

This study investigated the effect of cognitive load on the probability that participants derived the scalar inferences of seven scalar words: 'low', 'scarce', 'might', 'or', 'some', 'most', and 'try'.

Exp. 1 manipulated cognitive load by asking participants to memorise simple or complex grid patterns while they compared sentences against pictures. At a first glance, the results were puzzling: participants were significantly less likely to respond pragmatically to sentences with the scalar words ‘or’, ‘some’, and ‘try’ when they had to memorise complex grids, but there was no effect of memory load for any of the other scalar words. Once we also included into our analysis the data reported by van Tiel *et al.* (2019: Exp. 2), however, a more consistent pattern emerged, with participants being increasingly less likely to derive the scalar inferences of the positively scalar words ‘might’, ‘or’, ‘some’, ‘most’, and ‘try’—but not the negatively scalar words ‘low’ and ‘scarce’—when they were under greater cognitive load.

These data taken together resolve the shortcomings from van Tiel *et al.* (2019) which we mentioned above. The minimal- and maximal-load matrices lead to a clearer differentiation of the cognitive load effect, and ‘try’ now reliably patterns with the other positively scalar words, whose behaviour differs significantly from that of the negatively scalar words.

From a methodological perspective, the results of Exp. 1 show that the effect of memory load on the probability of scalar inferencing is difficult to detect. Thus, neither Exp. 1 nor van Tiel *et al.*’s (2019) Exp. 2 yielded a consistent set of results. Only when the results of the two experiments were put together did a consistent pattern of results emerge.

The volatility of the effect of memory load does not seem to be due to a lack of power: De Neys and Schaeken (2007) tested 56 participants; Marty and Chemla (2013) 16 participants; Marty *et al.* (2013) 26 participants, and Dieussaert *et al.* (2011) 106 participants—our Exp. 1 tested 100 participants. One notable departure from previous studies is that we tested memory load between participants rather than within. However, from a theoretical perspective, this difference should not influence the results (*cf.* also Charness *et al.*, 2012).

Exp. 2 manipulated cognitive load by varying the amount of time and effort participants could invest in the sentence-picture verification task. In the fast version of the experiment, participants saw the picture for a brief amount of time, whereupon it was replaced by the sentence; in the normal condition, the picture remained on screen when the sentence was presented; in the slow condition, participants were forced to wait for three seconds before providing their truth judgements. Unlike the results reported by Chevallier *et al.* (2008), we failed to find significant effects of processing time on the probability of deriving the scalar inferences of any of the seven scalar words.

Again, the failure to replicate Chevallier *et al.*’s results cannot be attributed to a lack of power: Chevallier *et al.* tested 59 participants, whereas we tested 150. In both cases, the available processing time was manipulated between participants.

Further research should determine whether the failure to replicate Chevallier *et al.*’s results is incidental or whether the available processing time really does not influence the rates of scalar inferences. More generally, also given the unstable results from the memory load experiments, it may be worthwhile to investigate the robustness of results from the experimental pragmatics literature in a more systematic and comprehensive way, as has been already done in various other fields (*e.g.* Cova *et al.*, 2018; Open Science Collaboration, 2015).

Such a large-scale replication should also take into consideration data from other experimental tasks, such as eye-tracking (*e.g.*, Huang and Snedeker, 2009), reading times (*e.g.*, Politzer-

Ahles and Husband, 2018), and ERP (e.g., Barbet and Thierry, 2018). Results from these experimental tasks seem more equivocal about the presence or absence of a processing cost for scalar inferencing, when compared to the results from sentence-picture verification tasks.

From a theoretical perspective, the results of Exp. 1 are in line with the view espoused by van Tiel et al. (2019) that the derivation of scalar inferences is not cognitively demanding per se, but only if the scalar inference introduces a negative proposition into the meaning of the sentence, as is the case for positively scalar words.

It has been shown that sentences that express negative information, even implicitly, are cognitively (e.g., Clark, 1970), semantically (e.g., Bierwisch, 1967), and syntactically (e.g., Heim, 2006) marked relative to sentences that do not. The late acquisition of negative scalar adjectives, such as ‘short’ and ‘low’, compared to their positive scalar counterparts ‘tall’ and ‘high’, also confirms the difficulty that underlies the processing of negative information (e.g., Klatzky et al., 1973).

A further question is *why* the processing of negative information should be cognitively costly. At least two possible answers to this question can be distinguished. One possibility is that hearers evaluate negative sentences by first evaluating their positive counterparts and then reversing the truth value, whereas positive sentences are evaluated directly (e.g., Clark and Chase, 1973). A second possibility is that negative sentences presuppose an expectation that their positive counterparts are true. The accommodation of this presupposition may be what makes the processing of negative information cognitively demanding (e.g., Moxey, 2006).

In either case, it seems plausible to suppose that the alleged processing cost of scalar inferences is in fact an idiosyncrasy due to the fact that most research has hitherto been concerned with positively scalar words, such as ‘some’ and ‘or’, rather than with negatively scalar words (but cf. Cremers and Chemla, 2014; Romoli and Schwarz, 2015).

It follows from the scalarity-based explanation that the apparent processing cost that has been observed for the scalar inferences of ‘some’ and ‘or’ should not be construed as evidence for the relevance-theoretic view that pragmatic inferencing is necessarily cognitively effortful. At the same time, the experimental record also fails to corroborate the defaultist prediction that scalar inferences are derived automatically, and that it is their overturning that is cognitively effortful. Rather, it seems that the literal and pragmatic meanings of scalar words can be accessed in parallel, without an intrinsic processing cost for either interpretation.

Of course, in order to arrive at a more decisive verdict about the adequacy of the scalarity-based explanation, it will be necessary to extend the purview to a larger sample of scalar words and experimental tasks. We leave this enterprise to future research.

## References

- Barbet, C. and G. Thierry (2018). When *some* triggers a scalar inference out of the blue. An electrophysical study of a Stroop-like conflict elicited by single words. *Cognition* 177, 58–68.
- Bierwisch, M. (1967). Some semantic universals of German adjectivals. *Foundations of Language* 3, 255–256.

- Bott, L. and I. A. Noveck (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language* 51, 437–457.
- Charness, G., U. Gneezy, and M. A. Kuhn (2012). Experimental methods: between-subject and within-subject design. *Journal of Economic Behavior & Organization* 81, 1–8.
- Chevallier, C., I. A. Noveck, T. Nazir, L. Bott, V. Lanzetti, and D. Sperber (2008). Making disjunctions exclusive. *Quarterly Journal of Experimental Psychology* 61, 1741–1760.
- Chevallier, C., D. Wilson, F. Happé, and I. Noveck (2010). Scalar inferences in autism spectrum disorders. *Journal of Autism and Developmental Disorders* 40, 1104–1117.
- Clark, H. H. (1970). The primitive nature of children's relational concepts: a discussion of Donaldson & Wales. In J. R. Hayes (Ed.), *Cognition and the development of language*, pp. 269–278. New York, NY: Wiley.
- Clark, H. H. and W. G. Chase (1973). On the process of comparing sentences against pictures. *Cognitive Psychology* 3, 472–517.
- Cova, F., B. Strickland, A. Abatista, A. Allard, J. Andow, M. Attie, J. Beebe, R. Berniūnas, J. Boudesseul, M. Colombo, F. Cushman, R. Diaz, N. N'Djaye Nikolai van Dongen, V. Dranseika, B. D. Earp, A. G. Torres, I. Hannikainen, J. V. Hernández-Conde, W. Hu, F. Jaquet, K. Khalifa, H. Kim, M. Kneer, J. Knobe, M. Kurthy, A. Lantian, S.-y. Liao, E. Machery, T. Moerenhout, C. Mott, M. Phelan, J. Phillips, N. Rambharose, K. Reuter, F. Romero, P. Sousa, J. Sprenger, E. Thalabard, K. Tobia, H. Vicianá, D. Wilkenfeld, and X. Zhou (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*.
- Cremers, A. and E. Chemla (2014). Direct and indirect scalar implicatures share the same processing signature. In S. Pistoia Reda (Ed.), *Pragmatics, Semantics and the Case of Scalar Implicatures*, pp. 201–227. London, United Kingdom: Palgrave Macmillan.
- De Neys, W. and W. Schaeken (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology* 54, 128–133.
- Deschamps, I., G. Agman, Y. Loewenstein, and Y. Grodzinsky (2015). The processing of polar quantifiers, and numerosity perception. *Cognition* 143, 115–128.
- Dieussaert, K., S. Verkerk, E. Gillard, and W. Schaeken (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology* 64, 2352–2367.
- Gazdar, G. (1979). *Pragmatics: implicature, presupposition, and logical form*. New York, NY: Academic Press.
- Geurts, B., N. Katsos, C. Cummins, J. Moons, and L. Noordman (2010). Scalar quantifiers: logic, acquisition, and processing. *Language and Cognitive Processes* 25, 130–148.
- Heim, I. (2006). Little. In M. Gibson and J. Howell (Eds.), *Proceedings of Semantics and Linguistic Theory 14*. Ithaca, NY: Cornell University.
- Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Ph. D. thesis, University of California, Los Angeles.
- Horn, L. R. (1989). *A natural history of negation*. Chicago, IL: University of Chicago Press.
- Huang, Y. T. and J. Snedeker (2009). On-line interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology* 58, 376–415.
- Klatzky, R., E. V. Clark, and M. Macken (1973). Asymmetries in the acquisition of polar adjectives: linguistic or conceptual? *Journal of Experimental Child Psychology* 16, 32–46.
- Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational impli-*

- capture. Cambridge, MA: MIT Press.
- Marty, P. and E. Chemla (2013). Scalar implicatures: working memory and a comparison with only. *Frontiers in Psychology* 4, 1–12.
- Marty, P., E. Chemla, and B. Spector (2013). Interpreting numerals and scalar items under memory load. *Lingua* 133, 152–163.
- Matsumoto, Y. (1995). The conversational condition on Horn scales. *Linguistics and Philosophy* 18, 21–60.
- Moxey, L. (2006). Effects of what is expected on the focussing properties of quantifiers: a test of the presupposition-denial account. *Journal of Memory and Language* 55, 422–439.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349(6251), 1–10.
- Politzer-Ahles, S. and M. E. Husband (2018). Eye movement evidence for context-sensitive derivation of scalar inferences. *Collabra* 1, 1–13.
- Romoli, J. and F. Schwarz (2015). An experimental comparison between presuppositions and indirect scalar implicatures. In F. Schwarz (Ed.), *Experimental perspectives on presuppositions*, pp. 215–240. Cham, Germany: Springer.
- Sperber, D. and D. Wilson (1986). *Relevance: communication and cognition*. Oxford, United Kingdom: Blackwell.
- van Tiel, B., E. Pankratz, and C. Sun (2019). Scalar and scalarity: processing scalar inferences. *Journal of Memory and Language* 105, 93–107.