SAARLAND UNIVERSITY

Faculty of Philosophy
Department of Language Science and Technology
MASTER'S THESIS

A Tale of Scales, Polarity, and Individual Differences

Submitted by
Michael Vrazitulis
Saarbrücken
25 May 2023

Supervisors:

Prof. Dr. Vera Demberg

Dr. Jia Ern Loy

Saarland University
Faculty of Philosophy
Department of Language Science and Technology
Campus – Building C7.2
66123 Saarbrücken
Germany

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged. I assure that the electronic version is identical in content to the printed version of the master's thesis.

M. Vrazitulis

Michael Vrazitulis

Saarbrücken, 25 May 2023

Acknowledgements

Special shout-outs to my little sister Kathi who drew the pictures for the sentence–picture verification task for me. (Thank God, because I cannot draw at all myself.)

I am very grateful to my thesis supervisors for their patience and their valuable guidance throughout this project.

The experimental studies reported in this thesis would not have been possible without funding from the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 94887).

erc
European Research Council
Established by the European Commission

Abstract

Recent studies by van Tiel et al. (2019) and van Tiel and 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 two experiments through online crowd-sourcing ($N = 100$ and $N = 400$), employing a sentence–picture verification task, we obtain mild support for this view, although large between-scale differences remain unaccounted for. In addition, we tested if individual differences in working memory capacity, print exposure, or fluid intelligence would further modulate response behaviour during SI processing. Most saliently, subjects with higher fluid intelligence appeared to slow down more drastically whenever encountering pragmatic ambiguity due to a potential SI.

Contents

Chapter 1
Introduction

Scalar implicature (SI) is surely one of the most extensively studied phenomena in the field of experimental pragmatics. So, one may rightfully wonder why it would be interesting at all to embark on yet another journey on that well-trodden path. Well, although it may not seem that way at first glance, there are still a lot of questions within that branch of inquiry that lack decisive answers. In fact, navigating through the jungle of competing findings, frameworks, and accounts on how humans process and produce SIs can be confusing at times.

Throughout the present work, our focus lies on the comprehension, i.e., processing aspect of SIs. To be more specific, our main goal is to put a recently emerged theory, first expressed by van Tiel et al. (2019), to the test. According to that theory, the cognitive effort associated with SI processing crucially depends on the polarity of the underlying scalar term: As an example, relatively high effort is predicted for processing the positively polar scalar term 'some' as implicating 'some, *but not all*', but relatively low effort for processing the negatively polar scalar term 'not all' as implicating 'not all, *although some*'. Claims about such differences in processing effort are then typically assessed through response times and response choices in experimental paradigms where participants are asked to judge the veracity of sentences that, by their nature, may elicit a psycholinguistic process of interest—here, the computation of SIs (or lack thereof).

Beyond an attempt at conceptual replication of effects predicted by the polarity hypothesis, the present work additionally focuses on assessing potential relationships between SI processing and individual differences among comprehenders in terms of working memory capacity, fluid intelligence, or print exposure.

In order to address the stated issues of interest, we conducted two experimental studies via online crowd-sourcing. The first of these studies ($N = 100$) did not yield evidence in favour of the polarity hypothesis. It was therefore followed up by a second study with a larger number of subjects ($N = 400$) and, thus, higher statistical power. That second study did, indeed, provide some support for the polarity hypothesis. Yet, closer examination of the results under different paradigms of statistical analysis revealed said support to be rather mild and, crucially, to be driven only by a minority of the assessed linguistic materials.

As for individual differences, these were only assessed in the first study ($N = 100$). Interestingly, we obtain the quite salient (but exploratory) result that participants who scored higher on a Raven's Progressive Matrices task, operationalising fluid intelligence, would slow down more drastically, on average, whenever faced with pragmatic ambuigity in the SI-related main task.

The presentation of this work's contents is structured as follows: Right below in Chapter 2, we discuss some of the important theoretical background behind scalar implicature processing, the polarity hypothesis, and the three psychological constructs deemed relevant for our present individual-differences analyses. On the basis of that, Chapter 3 then contains a brief and precise statement of the present work's aim. Subsequently, an overview of the experimental materials and design utilised in our present two studies is provided in Chapter 4. Those two studies that we eventually carried out relying on these materials are described in Chapters 5 and 6, including details on the various employed statistical methods and the obtained results. A final discussion, putting our present findings into broader context, follows in Chapter 7 and eventually leads up to the conclusion statement given in Chapter 8. Some supplementary information and materials can be found in Appendices A, B, and C.

Chapter 2
Background

Some theoretical background on scalar implicatures and the notion of scalar polarity is provided right below in Section 2.1. Previous work on individual differences in working memory capacity, fluid intelligence, and print exposure as well as their potential relation to scalar implicature processing is summarised in Section 2.2.

2.1 Scales and Polarity

Here, we want to review certain underlying concepts that motivate our main research question about scalar polarity.

2.1.1 Horn Scales

In past literature, there have been many different proposals of defining what constitutes a *Horn scale* (see, e.g., Horn, 1972; Gazdar, 1979; Atlas and Levinson, 1981; Hirschberg, 1991), which is an important concept to rely upon when talking about scalar implicatures. In particular, given some natural language, what are the criteria to decide if an arbitrary tuple $\langle e_1, e_2, \ldots, e_n \rangle$ of expressions drawn from that language constitutes such a scale or not? For our purposes, a satisfactory definition should be able, at least, to correctly identify the six English-language Horn scales ⟨some, all⟩, ⟨or, and⟩, ⟨not all, no⟩, ⟨not all, none⟩, ⟨scarce, absent⟩, and ⟨three, four, five⟩ as such, but to reject the following six ill-formed pseudo-scales: *⟨green⟩, *⟨is a, forgotten⟩, *⟨and, or⟩, *⟨not all, big⟩, *⟨exactly four, exactly five⟩, and *⟨some, some but not all⟩. Here, we attempt such a definition (semi-)formally, based on the following prerequisite ideas:

- $Worlds(S)$: The set of all worlds in which a well-formed sentence S is semantically true. E.g., our real world as of 2023 is an element of $Worlds($‘There are at least 8 billion humans.’$)$, whereas some counterfactual world may rather be an element of $Worlds($‘There are fewer than 8 billion humans.’$)$.
- $Worlds(\neg S)$: The set of worlds in which a well-formed sentence S is semantically false. E.g., our real world is an element of $Worlds(\neg$‘There are fewer than 8 billion

humans.'), but some counterfactual world may be an element of $Worlds(\neg$'There are at least 8 billion humans.') instead.

- $d_S(w_1, w_2)$: The (signed) degree of difference between two worlds w_1, w_2 along the dimension of meaning captured by a well-formed sentence S; in case that $w_1 = w_2$ then trivially $d_S(w_1, w_2) = 0$, otherwise $d_S(w_1, w_2) \neq 0$, and in particular $d_S(w_1, w_2) = -d_S(w_2, w_1)$. This function is hard to define in exact quantitative terms, but relying on an intuition of what comparisons some quantitative understanding of it may plausibly yield across non-trivial cases will be necessary for the subsequent definition of Horn scales to work. E.g., let us label our real world as w_{real}, a counterfactual world where there are only exactly 5 billion humans as $w_{5\,\text{billion}}$, and another counterfactual world in which there are, in fact, exactly 100 billion humans overall as $w_{100\,\text{billion}}$. Then, with regard to any well-formed sentence S that constitutes a statement about the number of humans in the world, we can plausibly assume that $|d_S(w_{\text{real}}, w_{5\,\text{billion}})| < |d_S(w_{\text{real}}, w_{100\,\text{billion}})|$ and also that $d_S(w_{\text{real}}, w_{5\,\text{billion}}) < 0 < d_S(w_{\text{real}}, w_{100\,\text{billion}})$. As an additional example, consider our real world w_{real} again where the freezing point of water is at 0 °C. Let us conceive of a counterfactual world $w_{-1\,^\circ\text{C}}$ in which the freezing point of water is instead at −1 °C. And further, in another counterfactual world, say $w_{1\,^\circ\text{C}}$, the freezing point of water shall lie at 1 °C. Now, with respect to any well-formed sentence S' that constitutes a statement about what the freezing point of water is, one may plausibly suppose that $|d_{S'}(w_{\text{real}}, w_{-1\,^\circ\text{C}})| = |d_{S'}(w_{\text{real}}, w_{1\,^\circ\text{C}})|$, but that $d_{S'}(w_{\text{real}}, w_{-1\,^\circ\text{C}}) < 0 < d_{S'}(w_{\text{real}}, w_{1\,^\circ\text{C}})$.
- $S[\alpha/\beta]$: A sentence that is derived by substituting the first occurrence of the expression α within the well-formed sentence S by the expression β. Note that $S[\alpha/\beta]$ may or may not be well-formed itself.

Now, we can define a tuple $\langle e_1, e_2, \dots, e_n\rangle$ of n linguistic expressions to be a **Horn scale** if and only if both of the following conditions 1 and 2 are met:

1. It is the case that $n \geq 2$.
2. There is a well-formed sentence S that contains e_n where,
   (a) for each $i \in \mathbb{N}$ with $1 \leq i < n$,
      i. the sentence $S[e_n/e_i]$ is well-formed
      ii. and $\varnothing \subset Worlds(S[e_n/e_{i+1}]) \subset Worlds(S[e_n/e_i])$,
   (b) and for no world $w \in Worlds(S)$ there are two worlds $w', w'' \in Worlds(\neg S)$ such that $d_S(w, w') < 0 < d_S(w, w'')$.

Note that subcondition 2-(a)-i ensures that all expressions fulfil a similar grammatical role, subcondition 2-(a)-ii ensures that the expressions are ordered by their degree of semantic informativeness, and subcondition 2-(b) ensures (perhaps least obviously) that the expressions are either purely lower-bounded or purely upper-bounded, i.e., display uniform polarity.

To illustrate the functioning of the provided conditions, let us first take a look at several instructive examples of tuples of English expressions that violate some of these conditions and consequently cannot be regarded Horn scales:

- Violation of condition 1: *⟨green⟩ does not fulfil condition 1 ($n = 1 < 2$). Therefore, it is not a Horn scale.

- Violation of subcondition 2-(a)-i: *$\langle$is a, forgotten$\rangle$ fulfils condition 1 ($n = 2 \geq 2$), but not condition 2 as there is no well-formed sentence S that contains 'forgotten' **where (a) [i]** S[forgotten/is a] is also well-formed.[1] Thus, it is not a Horn scale.

- Violation of subcondition 2-(a)-ii: *$\langle$and, or$\rangle$ fulfils condition 1 ($n = 2 \geq 2$), but not condition 2 as there is **no well-formed sentence** S that contains 'or' **where (a) [i]** S[or/and] is also well-formed **and [ii]** there are (conceivable) worlds in which S is true and all such worlds are bound to be worlds where S[or/and] is also true. In other words, 'or' does not entail 'and' (rather the other way around). Therefore, the given tuple is not a Horn scale.

- Violation of subcondition 2-(a)-ii: *$\langle$not all, big$\rangle$ fulfils condition 1 ($n = 2 \geq 2$), but not condition 2 as there is **no well-formed sentence** S that contains 'big' **where (a) [i]** S[big/not all] is also well-formed **and [ii]** there are (imaginable) worlds in which S is true and all such worlds have to be worlds in which S[big/not all] is true as well. In other words, 'not all' does not entail 'big' (e.g., 'Not all animals are horses.' does not entail 'Big animals are horses.'). Thus, the tuple in question is not a Horn scale.

- Violation of subcondition 2-(a)-ii: *$\langle$exactly four, exactly five$\rangle$ fulfils condition 1 ($n = 2 \geq 2$), but not condition 2 as there is **no well-formed sentence** S that contains 'exactly five' **where (a) [i]** S[exactly five/exactly four] is well-formed too **and [ii]** there are (imaginable) worlds in which S is true and all such worlds must be worlds in which S[exactly five/exactly four] is also true. In other words, 'exactly five' does not entail 'exactly four' (even though plain 'five' would, in fact, entail 'four' in most common sentences). Hence, the tuple given above is not a Horn scale.

- Violation of subcondition 2-(b): *$\langle$some, some but not all$\rangle$ fulfils condition 1 (since $n = 2 \geq 2$), but not condition 2 as there is **no well-formed sentence** S that contains 'some but not all' **where (a) [i]** S[some but not all/some] is well-formed as well **and [ii]** there are (conceivable) worlds in which S is true and all such worlds are bound to be worlds where S[some but not all/some] is also true, but not vice-versa, **and (b)** for no world w where S is true there are two worlds w', w'' where S is false that differ from w each from an opposite direction on the meaning dimension spanned by S (because, actually, there always are—e.g., for a world w where it is true that $S =$ 'Some but not all dogs are happy.' one can think of a world w' where one dog is happy and of a world w'' where every dog is happy, thus falsifying S from opposite directions on the meaning dimension of the quantity of existing happy dogs).[2] Therefore, the given tuple does not qualify as a Horn scale.

By contrast, let us now look at some examples of tuples that are, in fact, Horn scales:

- $\langle$some, all$\rangle$ fulfils condition 1 ($n = 2 \geq 2$) and also condition 2 as there is **a well-formed sentence** $S =$ 'Some dogs are happy.' that contains 'all' **where (a) [i]** S[all/some] = 'Some dogs are happy.' is also well-formed **and [ii]** there are (imaginable) worlds in which all dogs are happy and all such worlds must inevitably also be worlds in which some dogs are happy, but not vice-versa, **and (b)** for no world w in which all dogs are happy one can think of two distinct worlds w', w'' in which

[1] Okay, at least, I personally could not come up with any, but perhaps there is such a well-formed sentence somewhere in English after all.

[2] The theoretical difference between a tuple like $\langle$some, all$\rangle$ and a tuple like *$\langle$some, some but not all$\rangle$ as well as the reason why only the former should be considered a Horn scale—despite both of them being tuples of increasingly informative expressions—has also been addressed in terms of what is called the *symmetry problem* in earlier work, for instance, by Fox and Katzir (2011).

not all dogs are happy that differ from w each from an opposite direction on the meaning dimension of the quantity of existing happy dogs. In consequence, it is a Horn scale.

- $\langle$or, and$\rangle$ fulfils condition 1 ($n = 2 \geq 2$) and also condition 2 considering that there is **a well-formed sentence** $S =$ 'Mary is tall and poor.' that contains 'and' where **(a)** **[i]** S[and/or] = 'Mary is tall or poor.' is also well-formed **and** **[ii]** there are (imaginable) worlds in which Mary is tall and poor and all such worlds also have to be worlds in which Mary is tall or poor, but not vice-versa, **and (b)** for no world w in which Mary is tall and poor one may imagine two distinct worlds w', w'' in which Mary is not both tall and poor that differ from w each from an opposite direction on the meaning dimension of the strength of the logical connective linking Mary's potential state of being tall with that of her being poor. Thus, it is a Horn scale.
- $\langle$not all, no$\rangle$ fulfils condition 1 ($n = 2 \geq 2$) and also condition 2 as there is **a well-formed sentence** $S =$ 'No dogs are happy.' that contains 'no' **where (a)** **[i]** S[no/not all] = 'Not all dogs are happy.' is also well-formed **and** **[ii]** there are (imaginable) worlds in which no dogs are happy and all such worlds must also be worlds in which not all dogs are happy, but not vice-versa, **and (b)** for no world w where no dogs are happy there are two conceivable worlds w', w'' in which some dogs are happy that differ from w in opposite directions on the meaning dimension of the quantity of existing happy dogs. So, it is a Horn scale.
- $\langle$not all, none$\rangle$ fulfils condition 1 ($n = 2 \geq 2$) and also condition 2 since there is **a well-formed sentence** $S =$ 'None of the hot dogs have been eaten.' which contains 'none' **where (a)** **[i]** S[none/not all] = 'Not all of the hot dogs have been eaten.' is well-formed as well **and** **[ii]** there are (conceivable) worlds in which none of said hot dogs have been eaten and all such worlds are bound to also be worlds in which not all of said hot dogs have been eaten, but not vice-versa, **and (b)** for no world w in which none of said hot dogs have been eaten it is possible to think of two worlds w', w'' in which some of said hot dogs have been eaten that differ from w each from an opposite direction on the meaning dimension of the quantity of hot dogs in question that have been eaten. Therefore, it is a Horn scale.
- $\langle$scarce, absent$\rangle$ fulfils condition 1 ($n = 2 \geq 2$) and also condition 2 as there is **a well-formed sentence** $S =$ 'Vegetation on Mars is absent.' that contains 'absent' where **(a)** **[i]** S[absent/scarce] = 'Vegetation on Mars is scarce.' is also well-formed **and** **[ii]** there are (imaginable) worlds where vegetation on Mars is absent and all such worlds must also be worlds where vegetation on Mars is scarce, but not vice-versa, **and (b)** for no world w in which vegetation on Mars is absent one can think of two worlds w', w'' where vegetation on Mars is actually present that differ from w each from an opposite direction on the meaning dimension of the amount of existing vegetation on Mars. Hence, it is a Horn scale.
- $\langle$three, four, five$\rangle$ fulfils condition 1 ($n = 3 \geq 2$) and also condition 2 as there is **a well-formed sentence** $S =$ 'Ann has survived five wars.' containing 'five' **where (a)** **[i]** both S[five/three] = 'Ann has survived three wars.' and S[five/four] = 'Ann has survived four wars.' are also well-formed **and** **[ii],** on the one hand, there are (conceivable) worlds in which Ann has survived five wars and all such worlds must also be worlds in which Ann has survived four wars, but not vice-versa, and, on the other hand, there are (conceivable) worlds in which Ann has survived four wars and all such worlds must also be worlds in which Ann has survived three wars, but not vice-versa, **and (b)** for no world w where Ann has survived five wars one can

think of two worlds w', w'' where Ann has not survived five wars that differ from w each from an opposite direction on the meaning dimension of the number of wars that Ann has survived. Consequently, it is a Horn scale.[3]

2.1.2 Scalar Implicature Processing

Having established what a Horn scale is, we will now see how that formal concept can be used in order to model the pragmatic phenomenon of *scalar implicature:* A statement like the following,

(1) Some of my co-workers are male.

has two possible interpretations: For the given example, its *literal* interpretation suggests that *some, and possibly all* members of the group of co-workers being referred to are male, whereas its *pragmatic* interpretation proposes that *some, but not all* members of said group are male. The essence of this ambiguity can be easily understood once conceptualising the word 'some' as the weaker term of the Horn scale ⟨some, all⟩: Utterances employing a weaker term (= semantically less informative expression) of some Horn scale, e.g., 'some' with respect to ⟨some, all⟩, are often understood as excluding the possibility that a modified utterance containing a stronger term instead also holds, due to what, following Grice (1975), can be called the *maxim of quantity* that is required by the *cooperative principle* of communication. Such an implicit dismissal of the possibility that a more informative, alternative utterance could be true as well is precisely what is commonly referred to as scalar implicature (for short, SI). Crucially, it is the reason why listeners or readers who know a stronger alternative to the statement (1), here,

(2) All of my co-workers are male.

to actually be true may sometimes be inclined to judge (1) to be false, based on its pragmatic interpretation. Yet, other times, under the same circumstances, comprehenders may nevertheless judge (1) to be true, based on its literal interpretation, i.e., by ignoring the SI ('some' implies *not* 'all') that can be drawn here.

Regarding the psycholinguistic processing of underinformative statements (i.e., that employ a non-maximally informative scalar word) like (1), two opposing theories have been proposed: According to the *neo-Gricean* view (e.g., Levinson, 2000), comprehenders construct the pragmatic interpretation (i.e., draw the SI) by default and with low cognitive effort, but are able to opt out of it and instead adopt the literal interpretation once supportive context knowledge like (2) is considered, yet at the cost of high cognitive effort. By contrast, proponents of *relevance theory* (e.g., Sperber and Wilson, 1986) claim that it is the literal interpretation that is constructed by default and with low effort, whereas the pragmatic interpretation only arises, effortfully, if the comprehender deems the literal one not to be relevant enough within the specific context of conversation.

In their seminal study, Bott and Noveck (2004) conducted several experiments showing that, with regard to the ⟨some, all⟩ scale, participants responded significantly slower 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' on unambiguous sentences used as control trials. This finding came to be known as the *B&N effect.* Regarding the same Horn scale, De Neys and Schaeken (2007) found that participants who had to perform a dual task, consisting of (1) sentence verification and (2) memorisation of complex dot patterns, ended up verifying

[3] But note that the idea that numerals can form Horn scales is sometimes contested (see, for example, Spector, 2013).

underinformative sentences literally (rather than pragmatically) more often than control subjects who 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, but incompatible with the neo-Gricean view. That is, the effects suggest that drawing SIs is additionally effortful compared to interpreting underinformative sentences literally.

2.1.3 The Polarity Hypothesis

Rather recently, a further theory has emerged (van Tiel et al., 2019; van Tiel and Pankratz, 2021), claiming that the B&N effect—and, by extent, perhaps 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. Polarity, here, is defined as whether the literal meaning of a scalar term introduces a *lower bound*, for instance, on the number of co-workers in (1) or (2), rendering it positive, or if it rather introduces an *upper bound*, e.g., on the number of co-workers in (3) or (4) right below, rendering it negative.

(3) Not all of my co-workers are male.

(4) None of my co-workers are male.

Apart from their own studies' findings, van Tiel et al. point towards additional evidence that can be interpreted, partially, in support of this *polarity hypothesis* (evidence from: Cremers and Chemla, 2014; Romoli and Schwarz, 2015; Marty et al., 2020). Their theoretical explanation of why pragmatic interpretations arising from negative scales are computed equally as fast as or even quicker 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 and 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 and Pankratz (2021) theorise that a positively polar surface utterance like (1) evokes a negatively polar implicature once interpreted pragmatically, as shown in (1')—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 like (3), cognitively focusing on the literal meaning rather than on the implicature given in (3'), the latter being positively polar here, is more effortful (⇒ absent or reversed B&N effect).

(1') $[\text{Some}]_{+}$ of my co-workers are male. ⇝ $[\text{Not all}]_{-}$ (...)

(3') $[\text{Not all}]_{-}$ of my co-workers are male. ⇝ $[\text{Some}]_{+}$ (...)

2.1.4 A Secondary Hypothesis

Additionally, although this is not prominent within the scope of van Tiel and colleagues' line of research, a similar reasoning may lead us to the following prediction about a main effect of polarity on the verification response itself (i.e., whether, in a typical verification-task experiment, a subject classifies an underinformative statement as 'True' or as 'False'): Generally, when faced with two complementary alternatives of interpreting a statement, people might prefer choosing the one that they anticipate to be cognitively less effortful.

Therefore, one may expect to find a higher preference of the literal interpretation in cases like (1)/(1') than in cases like (3)/(3').

Only after conducting our present experimental studies, we actually stumbled upon previous work whose results could have further motivated this secondary hypothesis beyond the merely theoretical reasoning just laid out: Gotzner et al. (2018) report to have found the polarity of adjectival scales (e.g., positive ⟨adequate, good⟩ versus, e.g., negative ⟨low, depleted⟩) to be, indeed, a relevant predictor of how likely a subject will draw an SI, with higher SI rates being associated with negative scales.

2.1.5 Shortcomings 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 actually not that large and decisive yet.

The seminal study by van Tiel et al. (2019) that the claim ultimately originates from had set out to experimentally compare the properties of seven different Horn scales of which two, ⟨low, empty⟩ and ⟨scarce, absent⟩, happened to be the only negatively polar ones. Incidentally, however, these two scales were also the only adjectival scales that were included—the remaining five were based on verbs, pronouns, or conjunctions. Only due to the fact that ⟨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 first 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, van Tiel and Pankratz (2021)'s follow-up study 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 among each other. Yet, it displays other properties that may be viewed as shortcomings: First, its experimental design *does not* make direct comparisons within polarity-contrastive pairs of Horn 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 positive adjectival scales (among them, e.g., ⟨content, happy⟩, ⟨warm, hot⟩, ⟨ajar, open⟩) as a whole against an arbitrary set of 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 (a paradigm described in more detail below in Section 4.1.1) 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 arbitrarily chosen to represent either positive or negative polarity. Second, as a further caveat, van Tiel and Pankratz actually operationalise polarity in a rather innovative way: Instead of treating it as a binary factor as usual, they come up with a sophisticated latent metric that is continuous and ranges, among their examined adjectival scales, from 2.18 (very positive) to -1.83 (very negative). 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. Although interesting as an idea, it is unclear what that forcefully 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). Going further, van Tiel and Pankratz' implicit assumption that the polarity hypothesis should be similarly valid for either of the two presented conceptions of 'polarity' actually weakens the theoretical justification of their claim that is grounded in the difficulty of negation processing: While it is plausible that scalar expressions like 'not all' (explicitly) or 'scarce' (implicitly) contain hard-to-process, negative information, the same is far less clear for a scalar like 'rainy', unless of course the very concept of negation itself is also understood to encompass a psychological extension that is based on emotional valence.

Speaking of negation and its explicit (e.g., in expressions '**not** above', '**not** many', '**in**frequently', '**un**abundant') versus implicit (e.g., in respective synonyms 'below', 'few', 'rarely', 'scarce') variants, as prominently distinguished by Fodor and Garrett (1975),[4] another interesting remark has to be made: Van Tiel and Pankratz (2021) theorise that Horn 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). 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 than what is the case for implicit negation. However, neither van Tiel et al. (2019) nor van Tiel and Pankratz (2021) actually make the effort to experimentally examine explicitly negative Horn 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 is not experimentally assessed there at all. 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 and Pankratz (2021) cite in support of their polarity-based view is one by Cremers and Chemla (2014): In Cremers and Chemla'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 Horn scale ⟨some, all⟩ or the negative one ⟨not all, no⟩[5] against their world knowledge. For instance, participants would be shown this underinformative sentence: 'Some elephants are mammals.' The pattern of results from that experiment,

[4] As has hopefully become apparent intuitively from the given examples, the distinction made by Fodor and 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 it 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}.

[5] 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 and Chemla (2014) as 'direct (SIs)' vs. 'indirect (SIs)' instead, following a terminological and conceptual convention introduced by Chierchia et al. (2004).

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 Chemla 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 and 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.[6] 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. But van Tiel and Pankratz do not even mention that second experiment by Cremers and Chemla while citing them, which seems a bit arbitrarily selective. This appears to be even more odd once it is considered that Cremers and Chemla actually give much more weight to that second experiment in the overall interpretation and discussion of their findings, literally leading them to give their publication the title 'Direct [=positive] and indirect [=negative] scalar implicatures share the same processing signature', which directly contradicts van Tiel and colleagues' polarity hypothesis in whose support it is nonetheless pointed towards by the latter authors. Of course, there may be good reasons to disagree with Cremers and Chemla on their relative dismissal of the results of their 'Experiment 1' and embracement of the results of their 'Experiment 2' instead. But not elaborating on the reasons to do so and simply not mentioning the second experiment by Cremers and Chemla and its opposing result pattern at all is a bit intransparent on the part of van Tiel and Pankratz.

Romoli and Schwarz (2015) conducted a similar verification experiment as Cremers and Chemla (2014) did, and they are also cited by van Tiel and 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 presuppositions (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 end up displaying a *reversed* B&N effect for negatively polar SIs, their experiment lacks a comparative condition where positively polar SIs are also tested (for which then a regular, non-reversed B&N effect would be expected), simply because their research question had a different focus. 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⟩, ⟨$_X$ or $_Y$, $_X$ and $_Y$⟩ vs. ⟨not both $_X$ and $_Y$, neither $_X$ nor $_Y$⟩,[7] ⟨possible, certain⟩ vs. ⟨not

[6] This instruction-based (rather than intuition-based) experimental paradigm for assessing on-line SI processing was first used by Bott and 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'.

[7] Here, we use a notation where the subscripted X and Y expressions are to be understood as placeholders for linguistic expressions that stay fixed across different examined scales; in this particular case, X and Y each represent a noun phrase ('the apple' and 'the pepper'). Although intuitively graspable,

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 evidently done in an attempt to replicate De Neys and Schaeken (2007)'s experimental design and, thus, detect potential (reversed) D&S effects (see again Section 2.1.2). Unlike what van Tiel and colleagues may have predicted in analogy to the response-time-related polarity hypothesis, i.e., now, 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, although not cited by van Tiel and Pankratz (2021), a study by Bill et al. (2018) can also be interpreted through the lens of the polarity hypothesis: In an experiment that 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, 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. Due to reasons laid out above, both of these properties would, however, be desirable for an experimental study that seeks to produce reliable evidence for or against what is claimed by the polarity hypothesis.

We will return to a more systematic evaluation as well as synthesis of prior evidence regarding the polarity hypothesis in form of a Bayesian meta-analysis further below in Section 6.2.2.

Another aspect that has been entirely neglected in the previous studies just reviewed here is accounting for individual differences across subjects. A review of other work (i.e., not particularly about polarity in SI processing) which may motivate examining such individual differences in the present context of research as well is provided right below in Section 2.2.

this notation effectively extends our (semi-)formal definition of a Horn scale given back in Section 2.1.1 by an additional possible abstraction. We will make use of a similar notation again further below in Section 4.1.1 when describing the linguistic materials constructed for our own experiments here.

2.2 Individual Differences

In many areas of experimental psycholinguistics, the importance of accounting for properties that gradually differ across human subjects and, in turn, modulate their language-related behaviour has gained remarkable attention. Therefore, it seems natural to ask if such properties also play a role in how different people handle the processing of SIs of varying polarity. Here, this is going to be examined with respect to the three psychological constructs working memory capacity, fluid intelligence, and print exposure.

2.2.1 Constructs and Measures

Humans possess a particular cognitive system that allows for short-term storage and manipulation of information. That system has a limited capacity, often referred to as *working memory capacity* (WMC). Despite being a rather general psychological construct, WMC has been shown to also be important in language-related behaviour specifically (e.g., in language comprehension; see Just and Carpenter, 1992). There are various established ways of assessing the extent of an individual's WMC. One common way is to administer some type of complex memory-span task (for an overview, see Conway et al., 2005), e.g., an operation span task (OSpan; introduced by Turner and Engle, 1989), which requires participants to remember sequentially presented letters (or words) while having to solve simple math equations at the same time.

Another important construct in cognitive psychology is *fluid intelligence*, which describes an individual's ability to perform such activities of abstract reasoning that do not or only minimally rely on previously experienced learning processes (in contrast to crystallised intelligence; see Cattell, 1963). A psychometric multiple-choice test originally developed by Raven (1938), Raven's Progressive Matrices (RPM), has been widely adopted and acknowledged as one possible measure of fluid intelligence. During any trial of this test, participants are presented an array of geometric pieces where one piece is missing and then asked to fill in the missing piece (through relational reasoning) by selecting from several provided options.

When it comes to inter-individual variation in language-related behaviour, it often makes sense to also consider constructs that are rather linguistically grounded themselves: One such construct along which humans can widely differ with respect to each other is their past experience of exposure to printed language (or, for short, *print exposure*). It can be assessed in various ways, for example, through self-report measures (e.g., Greaney, 1980) or recognition tests (introduced by Stanovich and West, 1989), probing the ability to correctly distinguish between real and fabricated names of authors, magazines, or books. The latter kind of measures, e.g., an author recognition test (ART), has the advantage of being more objective and, therefore, more reliable than the former in the sense that self-reports enable subjects to give a biased (often, socially desirable) response, for instance, by exaggerating the extent of their reading habits.

2.2.2 Role in Pragmatic Processing

It has been found that people with higher print exposure, as measured by an ART, are more likely to draw coherence-relation inferences when presented with an appropriate contextual signal (Scholman et al., 2020), whereas WMC, as measured by a reading span task, has not shown any similar explanatory power. Furthermore, subjects with higher print exposure tend to rely on implicit cause interpretation during ambiguous pronoun resolution and referential prediction (Johnson and Arnold, 2021). Both of these results indicate that higher print exposure is associated with increased sensitivity to pragmatic cues—at least at the discourse level and with regard to reading comprehension. However, a study by Ryzhova et al. (2023), which had set out to examine the relation between various measures of individual differences and pragmatic inferencing triggered by informationally redundant utterances, did not find any significant effect of print exposure, but did find one of non-verbal (a.k.a. fluid) intelligence, operationalised through Raven's Progressive Matrices, instead.

As far as SIs of the kind discussed in Section 2.1 are concerned, Yang et al. (2018) report that the type of question under discussion preceding an underinformative statement significantly interacts (on the type of verification response, i.e., whether someone responds with 'True' or 'False') with individual subjects' cognitive resources (i.e., a latent construct that includes WMC as measured by a count span task) and their socio-pragmatic abilities, but does not do so with regard to print exposure. Interestingly, Feeney et al. (2004) report a small, yet significantly higher tendency by subjects with larger WMC, as measured by a count span task, to respond literally to underinformative statements that are based on the ⟨some, all⟩ scale. However, a later study by Dieussaert et al. (2011) could not reproduce this main effect of WMC (here, measured with an operation span task) on type of verification response. Yet, what Dieuassert et al. did find was an interaction effect of WMC and cognitive load (in a dual-task experiment analogous to the one discussed in the seminal study by De Neys and Schaeken, 2007): Those participants with low WMC who had their cognitive system additionally charged by forced memorisation of complex dot patterns were less inclined to draw SIs and thus respond pragmatically to underinformative statements than participants with similarly low WMC that—in separate experimental conditions—were put under less or even no additional cognitive load.

Taken together, the discussed findings suggest that all three individual-differences constructs (WMC, fluid intelligence, print exposure) each modulate some aspects of some types of pragmatic processing. But regarding the processing of SIs in particular, the most relevant modulator should *a priori* be expected to be WMC, judging from what previous studies have reported.

2.2.3 Role in Negation Processing

As discussed in Section 2.1, the main theoretical explanation for effects of scalar polarity addresses the idea that implicatures drawn from positively polar scalars are negative, whereas implicatures drawn from negatively polar scalars are positive. Combined with the generally accepted notion of the effortfulness of negation processing, this seems to clarify why literal interpretations of underinformative statements are easier (i.e., faster and, perhaps, preferred) than pragmatic ones for positively polar scalars, but vice-versa for negatively polar scalars. Therefore, it is reasonable to assume that a cognitive measure that accounts for individual differences in how various people deal with negation processing (in general) may also be worth controlling for when assessing how people behave when processing underinformative statements that are based on Horn scales of either polarity.

A hint at how WMC could relate to negation processing is provided by the following finding by Deutsch et al. (2009): In their 'Experiment 3', participants had to perform an affective-priming task that was structured as follows: Participants were presented an expression that consisted of two words, the first one being a qualifier (as translated from German: 'a' or 'no'), the second one being a noun (e.g., 'party' or 'funeral'). Qualifiers could either be affirming ('a') or negating ('no'), while nouns could either have positive (‘party’) or negative (‘funeral’) valence. The presentation of each such expression was followed by the display of a (random) Chinese ideograph. The participants (who all did not know Chinese) were then asked to judge the visual pleasantness of the ideograph. As expected, participants would usually rate the pleasantness of an arbitrary ideograph higher if previously presented an overall positive (either ‘a’ + positive-valence noun or ‘no’ + negative-valence noun) expression. However, this effect was mitigated in a group of participants that were put into a dual-task condition, i.e., had to memorise an eight-digit number simultaneously (thus taxing their WMC). Those cognitively charged participants gave similarly low pleasantness ratings after being presented a negative-valence noun, regardless of whether it was preceded by an affirming or negating qualifier. Possibly, this indicates that processing negations is cognitively demanding in such a way that when available WMC is limited—either experimentally by a dual task or innately in an individual—such processing may be more difficult.

Attridge and Inglis (2014) use a variant of the RPM task and correlate the resulting scores yielded from participants of their experiment to those same participants’ performances on a conditional-inference task. One of their reported findings is that when faced with a modus-tollens inference of the scheme (not-$p \rightarrow q$), (not-q) $\implies$ (p), i.e., where the conclusion is positive, the higher a participant’s score on the RPM task, the more likely they were to (correctly) affirm the inference. Such a relationship could not be found, though, with regard to modus-tollens inferences of the more simple type ($p \rightarrow$ not-q), (q) $\implies$ (not-p), i.e., with fewer explicit negations in the premises to be processed. Hence, it appears that fluid intelligence (as quantified through RPM scores) facilitates negation-heavy logical reasoning.

So far, very little research on print exposure in the context of negation processing has been conducted. Staab et al. (2008) are a notable exception. In their ERP study, they investigate if semantically unexpected words appearing either in affirmative or negative sentences induce an N400 effect. In doing so, they also control for several individual-differences measures, among which one can find an ART. However, they do not find any significant modulation of N400 amplitudes by ART-operationalised print exposure.

Overall, there seems to be moderate evidence for a modulation of negation processing by WMC and fluid intelligence, but yet no evidence for such a modulation by print exposure.

2.2.4 Differences Across Literal Responders

Tavano and Kaiser (2010) raise the concern that verification response choices by human subjects in reaction to an underinformative sentence do not necessarily correspond to whether or not an SI was drawn, despite this being an often-made implicit assumption in many studies. Based on their own experiment’s findings, they further argue that even people who consistently respond literally may still be drawing an implicature, but consciously deciding not to respond accordingly due to ‘a decision about what type of response the experimenter expected, whether an underinformative description was a “good” one or not’ (ibid.). This interpretation is backed up by the fact that, in their experiment, response times between two groups of subjects, grouped by whether they tended to consistently respond literally or pragmatically, did not differ above chance level with respect to response time. Since almost all participants in the experiment were university students, whom Tavano and Kaiser suspect ‘to have high inference and language skills’ (ibid.), they further suggest that responding literally despite drawing an implicature might be associated with such competences regarding logical reasoning and language use. This idea seems relevant when thinking about potential individual differences attributable to print exposure, given that print exposure is a metric designed to capture particular skills related to (written) language use. In fact, an influential anthropological study by Scribner and Cole (1981) supports the claim that a higher degree of literacy facilitates the acquisition of skills in logical reasoning. But also fluid intelligence seems to be a likely candidate as a measure that may predict success in logical reasoning, given its role, e.g., in the study by Attridge and Inglis (2014) already described above in Section 2.2.3. Everything considered, one might expect individual differences in print exposure or fluid intelligence to divide subjects who tend to respond literally to underinformative sentences into two kinds: relatively fast responders with low print exposure or fluid intelligence who do not draw an implicature vs. rather slow responders with high print exposure or fluid intelligence who do draw an implicature, but decide to suppress it after conscious logical reasoning through which they nevertheless deem the literal response to be the correct one eventually.

Chapter 3
Aim of the Present Work

Here, we briefly describe the two main purposes of our present work.

3.1 Conceptual Replication of Polarity Effects

With previously expressed theories and reported findings regarding polarity effects on SI processing, summarised in Section 2.1, as a starting point, our present work aims to directly address the following two hypotheses through the administration of a sentence–picture verification task (for details on this task type, see Section 4.1.1):

(i) The polarity hypothesis (Section 2.1.3): 'False' (i.e., pragmatic) responses to underinformative statements employing positively polar scalars take longer (i.e., are more effortful) than 'True' (i.e., literal) responses to such statements. However, the exact opposite is the case for underinformative statements employing negatively polar scalars. Here, to 'take longer' does not refer to comparisons of raw response times, but to comparisons of response times from which any effects of verification response ('True'/'False') and polarity (positive/negative) occurring on *non*-underinformative statements have already been residualised out.

(ii) A secondary, polarity-based hypothesis (Section 2.1.4): 'False' (i.e., pragmatic) responses to underinformative statements employing positively polar scalars are less frequent (i.e., avoided due to anticipated effortfulness) than such responses to underinformative statements employing negatively polar scalars.

How these hypotheses are tested statistically is laid out further below in Section 5.3.4.

3.2 Probing the Role of Individual Differences

The first experimental study (Chapter 5) that is part of our present work does also feature exploratory tests designed to determine if and how individual differences in working memory capacity, fluid intelligence, or print exposure further modulate response times and response choices during SI processing. The process of statistical model selection along which these tests are performed is described in Section 5.4.2.

Chapter 4
Towards an Experiment

Just like several previous studies on SIs (e.g., Tavano and Kaiser, 2010; Marty et al., 2013; van Tiel et al., 2019; van Tiel and Pankratz, 2021), the two studies (Chapters 5, 6) conducted as part of our present work employ a sentence–picture verification paradigm in order to test the behaviour of human comprehenders when they are presented an underinformative statement that may or may not be interpreted with an SI.

Secondly, three tasks designed to measure individual differences in working memory capacity (operation span task), fluid intelligence (Raven's Progressive Matrices), and print exposure (author recognition test) are additionally administered to participants of our first study (Chapter 5).

4.1 Tasks

Find the materials chosen and created for the sentence–picture verification task in Section 4.1.1 as well as short discussions of the selected variants of the three tasks that elicit individual-differences measures in Sections 4.1.2, 4.1.3, and 4.1.4, respectively.

4.1.1 Sentence–Picture Verification

A sentence–picture verification (SPV) task consists of a series of pairs of a sentence and a picture (Clark and Chase, 1972). Usually, either the sentence is displayed before the picture, or both are displayed simultaneously. For each presented sentence–picture pair, a participant of the task has to give a binary response ('good description' or 'bad description', to be interpreted as 'True' or 'False' verification) depending on whether they consider the sentence to be a good or a bad description of the situation conveyed by the picture. Giving such a response is typically done by pressing either of two accordingly assigned buttons (or keys on a keyboard). Both the verification response itself as well as the time it took the participant to respond are relevant measures that can be collected from this task (and subsequently analysed).

Eight different Horn scales are used in the present experiments, as listed in Table 4.1. Four of these scales display positive polarity, and each positively polar scale can be con-

Table 4.1: Overview of the eight Horn scales which are probed in the present experiments. Each positively polar scale has a negatively polar counterpart. Note that, in the scales pertaining to the Time item, the subscripted expressions *VB* and *VBD* are placeholders that indicate the presence of a verb in either base form (e.g., ‘take’) or past-tense form (e.g., ‘took’), respectively. Analogously, in the scales for the Space item, the subscripted expression *PP-LOC* marks the presence of a locative prepositional phrase (e.g., ‘on the mountain’).

Item	Positively Polar	
Quantity	⟨some,	all⟩
Possibility	⟨might be,	is definitely⟩
Time	⟨sometimes $_{VBD}$,	always $_{VBD}$⟩
Space	⟨somewhere $_{PP\text{-}LOC}$ it is,	everywhere $_{PP\text{-}LOC}$ it is⟩
Item	Negatively Polar	
Quantity	⟨not all,	none⟩
Possibility	⟨might not be,	is definitely not⟩
Time	⟨did not always $_{VB}$,	never $_{VBD}$⟩
Space	⟨not everywhere $_{PP\text{-}LOC}$ is it,	nowhere $_{PP\text{-}LOC}$ is it⟩

Table 4.2: The four sentence items for the currently described sentence–picture verification task in all their possible variants.

Item	Scalar Type	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 4.3: 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 (**T** = 'True'; **F** = 'False'). Trial types *t1*+ (positively polar) and *t1*− (negatively polar) are the critical ones of interest as their underinformativeness reasonably allows for both 'True' (literal) and 'False' (pragmatic) responses.

Polarity	**Scalar Type**	**Picture Variant**		
		0 %	50 %	100 %
Positive	$\exists$	**F** *t3+*	**T** *t2+*	**T/F** *t1+*
	$\forall$	**F** *t6+*	**F** *t5+*	**T** *t4+*
Negative	$\neg\forall$	**T/F** *t1−*	**T** *t2−*	**F** *t3−*
	$\neg\exists$	**T** *t4−*	**F** *t5−*	**F** *t6−*

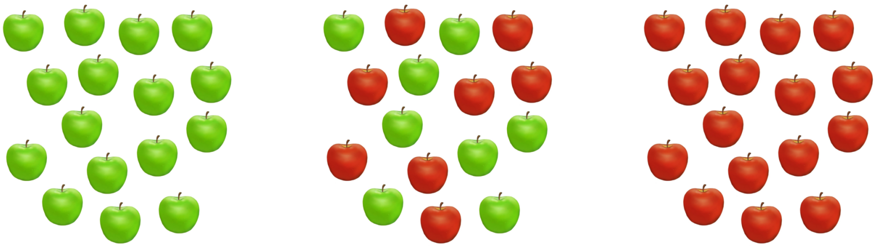


Figure 4.1: The three picture variants of the Quantity item corresponding to Horn scales ⟨some, all⟩ and ⟨not all, none⟩. The percentages annotated below indicate the gradual degree of informative fulfilment based on the positive scale.

trasted 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 (again, see Table 4.1). One thing that the scales included in our items all have in common is that they exhibit a *logical norm of correctness*.[8] That is, as opposed to what holds for other kinds of Horn 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 ⟨$\exists$, $\forall$⟩ (positive polarity) and ⟨$\neg\forall$, $\neg\exists$⟩ (negative polarity), respectively. To give two examples: The scalars 'some' and 'might be' behave like existential quantifiers ($\exists$) on the respective meaning dimensions of quantity (of something measureable) and possibility (of some event). Correspondingly, 'all' and 'is definitely' behave like universal quantifiers ($\forall$) on the same meaning dimensions. The weaker scalemates of the analogous negatively polar Horn scales, 'not all' and 'might not be',[9] behave like negated universal quantifiers ($\neg\forall$), while their stronger scalemates, 'none' and 'is definitely not',[10] behave like negated existentials ($\neg\exists$).

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

[9] 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.

[10] See footnote 9.

In the SPV task that is administered here, these scalar expressions appear within the sentences given in Table 4.2. These sentences, in turn, are associated with the pictures displayed in Figures 4.1, 4.2, 4.3, and 4.4. 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.

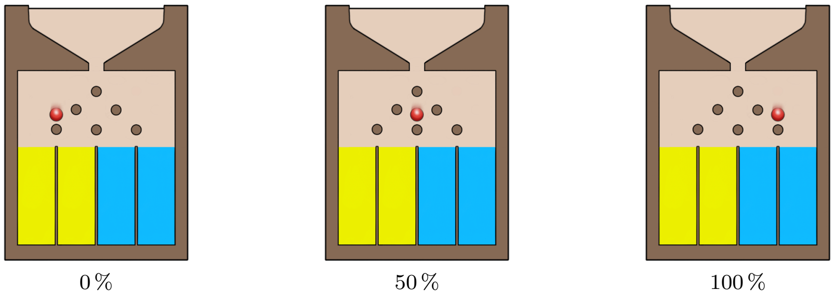


Figure 4.2: The three picture variants of the Possibility item corresponding to Horn scales ⟨might be, is definitely⟩ and ⟨might not be, is definitely not⟩. The percentages annotated below indicate the gradual degree of informative fulfilment based on the positive scale.

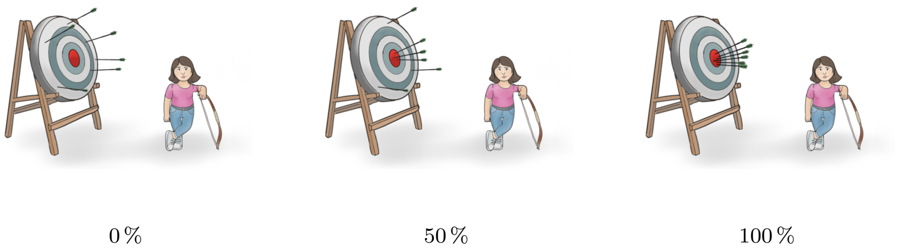
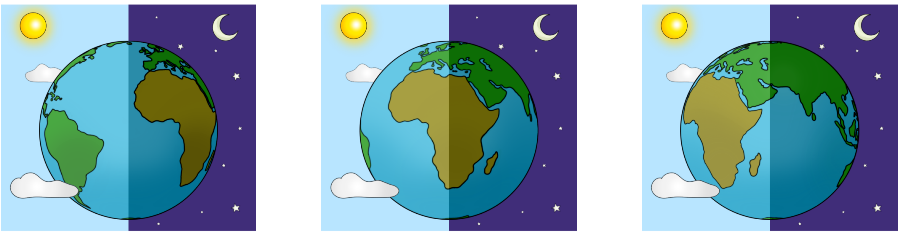


Figure 4.3: The three picture variants for the Time item corresponding to Horn scales ⟨sometimes *VBD*, always *VBD*⟩ and ⟨did not always *VB*, never *VBD*⟩. The percentages annotated below indicate the gradual degree of informative fulfilment based on the positive scale.

The picture type for the Quantity item displays 16 apples that can be either green or red (Figure 4.1). In the 0 % picture variant, every apple is green, whereas in the 100 % picture variant, every apple is red. In the 50 % picture variant, eight (i.e., half of the) apples are green, while the remaining eight (i.e., another half of the) apples are red. Note that the partitive number of apples that are red in the 50 % variant (8) compared to the overall number of apples (16) constitutes a scenario in which the use of a scalar like ‘some’ is typically perceived as natural and adequate (see Degen and Tanenhaus, 2015), unlike what would be the case for smaller numbers, where resorting to using a numeral as a descriptor instead would be perceived as more natural by human language users.

For the Possibility item, we opt for a picture type that shows a simple Galton board with four bins that a bead thrown into it might end up in. The two bins on the left are coloured in yellow, while the two bins on the right are coloured in blue. In the 0 % variant of that picture type, a single bead is displayed as being about to fall onto a round peg which will, in turn, direct it towards eventually falling into either of the two leftmost



0 % 50 % 100 %

Figure 4.4: The three picture variants of the Space item corresponding to Horn scales ⟨somewhere PP-LOC it is, everywhere PP-LOC it is⟩ and ⟨not everywhere PP-LOC is it, nowhere PP-LOC is it⟩. The percentages annotated below indicate the gradual degree of informative fulfilment based on the positive scale.

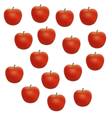
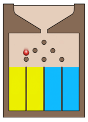
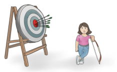
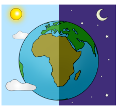
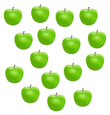
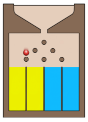
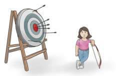
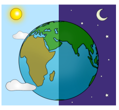
bins, both of which are yellow. In the 50 % variant, the bead can be seen to be about to fall onto a peg in the middle instead, which means that it can end up either in the second bin from the left (yellow) or in the second bin from the right (blue) with equal probability. Finally, in the 100 % variant, the bead is initially placed above a peg on the right side of the picture, now suggesting that it has to land in either of the two rightmost, blue bins.

The Time item's picture type is a rather intricate one: On the left side of any of its variants, an archery target with six arrows stuck into it is displayed. To the right of it, we see a lady leaning onto an archery bow, thus suggesting that she has used that bow to shoot said arrows onto the target next to her. What differs across picture variants, here, is the position of the arrows on the target: In the 100 % variant, all six arrows are located within the narrow bullseye of the target in its very center. The bullseye is clearly marked and set apart from the outer layers of the target by being coloured red. Yet, in the 50 % variant, only three of the six arrows have made it into the bullseye, while the remaining three are located somewhere on the outermost layer of the target. To complement this, the 0 % variant shows all six arrows placed on the outermost layer.

Lastly, the picture type for the Space item is designed as follows: It features an image of the Earth from such a perspective that the uppermost point on the visible hemisphere is the North Pole, the lowermost point the South Pole, and with the straight vertical line connecting these two points representing either a meridian slightly to the west of the westernmost point of Africa (0 % variant), slightly to the east of the easternmost point of Africa (100 % variant), or a meridian between these two that roughly splits Africa in half (50 % variant). Now, crucially, this picture type displays everything left to that vertical line in a lighter shade, but everything to the right of it in a darker shade. Additionally, the left half of the picture is extended by a cartoonish representation of daytime, using a light blue background and a symbolic image of the sun and two clouds, whereas the right picture half is extended by an analogous representation of nighttime, employing a dark violet background and a symbolic image of the moon and several stars.

In Table 4.3, 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 'True'

Table 4.4: 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 (**T** = 'True'; **F** = 'False') are listed.

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

(literally) or as ‘False’ (pragmatically). The remaining ten trial types are labelled *t2…t6*± (for details, again, Table 4.3) and constitute pragmatically unambiguous combinations. With every trial type being featured by four different sentence–picture items, i.e., the four items labelled Quantity, Possibility, Time, and Space, this gives rise to 48 potential 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 4.4.

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 can never appear immediately one after the other.

On a final note, now that we have already established the set of notations {Quantity, Possibility, Time, Space} to refer to our items, we want to introduce another set of notations, {+quantity, +possibility, +time, +space, −quantity, −possibility, −time, −space}, which we will use as a shorthand from now on to refer to the underlying individual scales (incl. associated sentences/pictures), half of which are positively polar (+) and another half of which are negatively polar (−).

4.1.2 Operation Span Task

In an operation span task (OSpan), participants are requested to judge the correctness of simple math equations—e.g., $(6 - 5) \times 2 = 5$—while simultaneously remembering an ordered set of letters (or words, symbols etc.) that incrementally increases in length. A new letter is introduced after each presentation of an equation. This testing process is repeated with multiple sets of varying final size. The mean proportion of letters in a set that is recalled correctly is a possible measure of interest, also known as the partial credit unit (PCU) score.

Here, an implementation of OSpan originally developed by Scholman et al. (2020), which can be administered via online crowd-sourcing, is used. However, because the original version of this implementation tends to display ceiling effects in participants’ performance (see Mayn and Demberg, 2022), a later adaptation (*v2_0622*) that has been slightly re-designed by Mayn in order to avoid such ceiling effects is opted for in the present experimental work. It encompasses 75 equations (of which 37 are correct) and its final set sizes range from three to seven, while there are three sets for each set size. The further set size two is used for practice trials. Sets are presented in random order, but that random order is always the same across participants.

4.1.3 Raven’s Progressive Matrices

Participants of a Raven’s Progressive Matrices (RPM) test are serially presented arrays (e.g., 3×3 ones) of geometric pieces from which one piece is always missing. On the presentation of each array, they are asked to select the correct missing piece out of a set of given options (e.g., eight options).

In the present work, we utilise the same online version of the test as Mayn and Demberg (2022) do, which consists of ten 3×3 arrays that are presented in order of increasing difficulty. The first two of these ten arrays (i.e., for the easiest trials) each come with six possible response options, while the remaining eight arrays are each presented together with eight possible response options (out of which only one is correct).

4.1.4 Author Recognition Test

Generally, an author recognition test (ART) involves presenting a series of potential names of famous literary authors to participants, with some of them belonging to actually existing and renowned authors, but the remaining ones being artificial. The measure of interest, assumed to operationalise a subject's print exposure, is how accurately they are able to tell real and fake author names apart.

In the two present studies, we use an automated version of that kind of test whose core has been implemented by Scholman et al. (2020) and which encompasses 130 potential author names, 65 of which belong to real authors, while the other 65 are fabricated. Here, the list of real author names is taken from Acheson et al. (2008), whereas the fake author names are adopted from Martin-Chang and Gould (2008). Examples of real names in the data set are *Isabel Allende*, *Maya Angelou*, or *Isaac Asimov*. Among the fake names, creations like *Lauren Adamson*, *Eric Amsel*, or *Carter Anvari* can be encountered.

4.2 Participant Sampling and Exclusion Criteria

People who are native English speakers, who are US citizens living in the US, who fall into the age range of 30–40 years (i.e., an 11-year time span)—and who have normal or corrected-to-normal vision while *not* being colour-blind[11]—are recruited as participants for our present experimental studies. The specific age range is motivated by the generational sensitivity of the author recognition test and adopted from Scholman et al. (2020), but increased by five additional years in order to capture roughly the same, recommended population of people in the present work, which is conducted five years later (Scholman et al.'s data is from 2018).

This is our main exclusion criterion for both experimental studies (Chapters 5 and 6) that we report here: In analogy to the experiment conducted by van Tiel and Pankratz (2021), participants who have responded erroneously to more than 20 % of control trials in the SPV task are excluded from any subsequent analyses.

The following two further exclusion criteria are applied only in our first study (Chapter 5) in which we assess individual differences as well: In line with Mayn and Demberg (2022), participants who have wrongfully judged the correctness of more than 20 % of math equations in the OSpan task are excluded. Finally, following practice advocated by Scholman et al. (2020), participants with a negative score in the author recognition test (see Section 5.4.1 for score calculation) are also excluded.

Given the above criteria for excluding participants, we sample until we reach $N = 100$ includible subjects in our first study (Chapter 5), but $N = 400$ includible subjects in our second study (Chapter 6). Additional explanations and justifications of these particular post-exclusion sample sizes are provided in Sections 5.2 and 6.2, respectively.

Further, in both of our presently reported experimental studies, sentence–picture verification responses with response times that are either shorter than 200 ms or at least as long as 10,000 ms are treated as missing data points. This is done in order to avoid treating accidental button presses ($<$ 200 ms case) or implausibly slow responses ($\geq$ 10,000 ms case) as valid observations that merit analysis.

[11] This sampling restriction based on vision capabilities is evidently necessary considering the nature of our sentence–picture verification items. Hence, it was applied as such in both of our present experimental studies. Note, however, that we failed to also explicitly mention this vision-based restriction in the corresponding preregistration protocols (**osf.io/pzja3**, **osf.io/dhpzq**) as would have been more precise.

4.3 Procedure

Both of our experiments (Chapters 5 and 6) were run via Prolific (**www.prolific.co**), relying on an implementation in the Lingoturk (Pusse et al., 2016) framework. They were hosted on a web server pertaining to Saarland University. Participants accessed either of the experiments using a common web browser.

It took roughly 30 minutes for a participant to complete the first experiment (Chapter 5), but only 6 minutes for a participant to complete the second experiment (Chapter 6) which consisted just of the SPV task. Accordingly, every participant in the first experiment was paid 5.29 GBP (≈ 6 EUR when conducted),[12] whereas every participant in the second experiment was paid 1.06 GBP (≈ 1.20 EUR when conducted), following our objective to compensate participants based on the legal minimum wage in Germany of 12 EUR/h as of the year 2023.

4.3.1 Completing the Tasks

Initially, participants would need to accept their participation in the respective study through a mouse click, resulting in the appearance of a first slide with instructions. These instructions inform about the estimated duration of completing the study, explain how to perform the SPV task, which is about to begin, and include a disclaimer pointing out how the collected data is going to be analysed in an anonymised fashion for ongoing research and that participation is voluntary.

After confirming through a mouse click that one has carefully read these instructions, a slide that asks for one's Prolific worker ID is displayed. Once that ID is provided, the sentence–picture verification task begins. It 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 description', to be interpreted as 'True' verification) or **0** (= 'bad description', to be interpreted as 'False' verification) is pressed, or once an automatic time-out of ten seconds is exceeded.[13]

In the case of our second experimental study (Chapter 6) where only the SPV task was administered, participants would now immediately be redirected to a final slide asking them to provide basic demographic information on themselves as well as to optionally share general feedback. That slide is described in the next section (Section 4.3.2).

Only in the case of our first experimental study (Chapter 5), the experiment would still continue after the SPV task as described in the following paragraphs:

[12] In the first experiment, there were eight subjects who only provided response data for the SPV task, but not for the subsequent three individual-differences tests. Although their data could therefore not be used in our main analyses, we did rely on it later, for a prospective power analysis in preparation to our second experimental study (see Section 6.2.1). Thus, we decided to compensate those eight anomalous subjects partially, each with 1.06 GBP (≈ 1.20 EUR), based on the approximate completion time of only the SPV task.

[13] Unfortunately, as we realised only after collecting all data, this time-out mechanism only worked when running the experiment in a Chromium-based web browser like Google Chrome, Opera or Microsoft Edge, but not in certain other browsers, e.g., Firefox, likely due to a JavaScript portability problem. This issue resulted in a very small minority of participants never experiencing such a time-out and thus being able to respond to individual trials even after up to 18 seconds, in the most extreme case. We dealt with such anomalous cases (only affecting 3 out of 4,800 trial observations in the first study as well as 9 out of 19,200 observations in the second study) by removing any such data points with ≥ 10 sec. response times prior to statistical analysis, as also specified in the discussion of exclusion criteria in Section 4.2. For more details on this issue, please consult the first page of this document: **osf.io/zrveq**.

After the SPV task, the OSpan task follows immediately, with an initial instruction slide. This slide announces a practice round during which only judgement of the correctness of math equations (but no letter memorisation) will have to be performed. This first practice round starts, upon again confirming one's Prolific worker ID, with the presentation of the first practice equation's (complex) left-hand side, with an equation sign and a question mark attached, e.g., '$(6 - 5) \times 2 = ?$'. Subsequently, once a button reading 'Verify' is clicked, the right-hand side of the equation (i.e., an alleged single-number result) is shown, for instance, '5'. To continue, participants have to click on either of two buttons corresponding to a 'correct' or 'incorrect' judgement. It is possible to click on a button only after a buffer time of one second from the onset of result display, though. Next, feedback on whether their judgement was right is provided, displayed for 800 ms. Subsequently, the next practice equation is presented, and this process is repeated eight times. Then, a second instruction slide is shown, announcing a second practice round where feedback slides are now replaced by slides displaying single letters that will have to be remembered in their order of appearance. After confirmation through a mouse click, participants are shown the first equation of the second practice round, succeeded—after eventually providing the judgement of its correctness—by the first letter that needs to be memorised. In the following, the next practice equation from the same set is displayed. After going through all equations and letters for this set, a slide follows which asks participants to type in the series of presented letters as best as they can recall it. Then, the second set is initialised. Once the two practice sets have been completed, a third instruction slide appears. It informs the participants that the real task is about to start and that it will be structurally equivalent to the second practice round, but with sets ranging from sizes three to seven. Upon confirmation, the real task begins, involving all 75 equations designated for it as trials, of which none have been seen previously in the practice rounds. During the real task, an automatic time-out individually set at the mean response time plus 2.5 standard deviations recorded in the first practice round limits the duration of equation display. Right after this task, a slide with two questions is shown which participants can optionally respond to by typing into text fields. The first question is: 'What strategy, if any, did you use for remembering the letters?' The second question reads: 'Did you write anything down during this experiment? If so, what did you write down? Please be honest; you will be paid in full regardless of your answer.'

Thereafter, a further instruction slide appears which informs about the author recognition test (ART). This slide instructs participants to answer if they recognise a presented name as one of a real author (by pressing **1** for 'Yes' or **0** for 'No'), but crucially not to resort to guessing as there would be a penalty for wrongfully classifying fake author names as real. Trials for this task are presented in alphabetical order (with regard to last names). The first trial is presented as soon as one's Prolific worker ID has been re-confirmed. Each recognition trial consists of (1) an empty slide displayed for 500 ms and (2) a slide showing a name that disappears once a recognition response is given (through pressing **1** or **0**), or once an automatic time-out of ten seconds is exceeded. Upon completion of all trials, a slide with two questions is displayed. The first question needs to be responded to by clicking on either of two radio buttons labelled 'Yes' and 'No', respectively, and reads: 'Have you seen these questions before?' The second question can be optionally responded to by typing into a text field; it reads as follows: 'Did you consult any external resources during this test? If so, which ones? Please be honest; you will be paid in full regardless of your answer.'

Eventually, instructions for the RPM test are displayed. The first out of a total of ten trials of the RPM test is shown after worker-ID re-confirmation. Each trial consists of a single slide that presents a visual array with a missing piece alongside either six or eight options provided for its completion. The preferred option can be selected by clicking on it. Once all trials have been administered, a slide with two questions is shown. The first one reads: ‘Have you seen these questions before?’ An answer to this question has to be provided by clicking on either of two radio buttons labelled ‘Yes’ and ‘No’, respectively. The second question can optionally be responded to by typing into a corresponding text field, and it reads: ‘If you selected “Yes” above: How long ago did you see them? Can you describe the kinds of patterns you have seen before (e.g., “the one with the wobbly lines”)?’

Now, having completed all four tasks, participants are directed to the final slide that asks for demographic information and general feedback, described right below in Section 4.3.2.

4.3.2 Demographics and Feedback

At the very end of either of our two experiments, a slide with several demographic questions as well as questions encouraging feedback remarks on the experiment as a whole is presented. These questions read and are structured as follows:

- ‘Gender’ (text field, obligatory)
- ‘Mother’s first language’ (text field, optional)
- ‘Father’s first language’ (text field, optional)
- ‘Are you bilingual (grown up with more than one language)?’ (radio buttons ‘Yes’/ ‘No’, obligatory)
- ‘Please list the languages you speak at an advanced level.’ (text field, optional)
- ‘Did you experience any technical difficulties or interruptions during the experiment?’ (text field, optional)
- ‘Do you have any other questions or comments for the researcher? If so, please type them here.’ (text field, optional)

Note that further demographic information about the subjects, concerning their age, biological sex, ethnicity, country of birth, and student/employment status, is already provided by Prolific itself by default and thus does not have to be queried explicitly.

4.4 Pre-Test and Pilot Studies

In December 2022, we created a pre-testing questionnaire aimed at assessing the validity of the linguistic and visual materials designed for the SPV task. This questionnaire was eventually answered by 20 native speakers of English who were informally sampled by contacting personal acquaintances. Each respondent had to go through 40 sentence–picture combinations that resembled the control trials of our present experiments, i.e., were intended to display no pragmatic ambiguity whatsoever. Similarly to our present experimental design, the order of presentation was individually randomised for each respondent. The collected pre-test responses were then analysed exploratorily in order to motivate final adjustments to our sentence–picture items presented in Section 4.1.1. That is, the pre-test itself actually featured prototypical (and somewhat imperfect) versions of these sentence–picture items that were a bit different from the final versions eventually opted for, here. For a more-in-depth discussion of that conducted pre-test, its results, and their implications, please have a look at Appendix A.

On 28 January 2023, an initial pilot study was run, which structurally resembled our first actual study (Chapter 5) in that it featured all four experimental tasks (SPV, then OSpan, then ART, then RPM), implemented within Lingoturk. It was administered to a sample of only five Prolific participants. Its purpose was to ensure the proper functionality of our experiment and to acquire a realistic estimate of average completion time. The response data from this initial pilot study can be accessed via the following link, in anonymised form: **osf.io/qr6bs**.

A second pilot study in which only the SPV task was administered was carried out on 13 February 2023 on a sample of 20 further Prolific participants. Here, the goal was to test if an adapted version of the instructions for the SPV task (now emphasising that people should strive to respond quickly and based on their first intuition) would elicit a higher ratio of pragmatic responses (now 20.8 %) than the rather low one found in the initial pilot study (only 10.0 %). Anonymised response data gathered from that second pilot study can be accessed here: **osf.io/ayj54**.

Eventually, a third pilot study, again featuring only the SPV task, was run on yet another sample of 20 subjects recruited from Prolific on 16 February 2023. It incorporated the adaptations made prior to the second pilot study, but had its framing of the instructions for the SPV task further modified as being about judging whether a sentence 'is a good or a bad description of the picture' (as finally also used in our two actual studies). This framing corresponds to common practice in earlier SPV-based studies probing SIs (e.g., van Tiel et al., 2019 or Tavano and Kaiser, 2010). By contrast, the versions of the task used in our first two pilot studies had been framed more directly as being about deciding if a sentence 'is true or false with regard to the picture', which by its more logical-sounding wording might have primed participants to pursue literal responses far more often. And indeed, results of the third pilot study show a much higher ratio of pragmatic responses (38.8 %) than in the pilot studies prior to it.[14] Here is a link to the anonymised response data collected from the third pilot study: **osf.io/mqpk2**.

None of the pre-test or pilot data is analysed with regard to our main research purpose, and none of the pilot studies' subjects were able to sit again for either of our two actual studies reported in Chapters 5 and 6.

[14] Of course, this observation may raise important theoretical questions about what it actually means to consider a sentence either 'true' or rather 'a good description' of some situation. Does either of those two instructional framings yield response data that is a cognitively more plausible proxy for the on-line processing behaviour in human comprehenders when they encounter pragmatic ambiguity? And why? We will come back to a quick discussion of these questions further below, in Section 7.4.

Chapter 5
Study One

This chapter is dedicated to describing the first study that we have conducted based on the experimental materials and design just presented in the previous Chapter 4. It was carried out using the crowd-sourcing platform Prolific on 6 and 7 March 2023. Hereinafter, we may also refer to this first experimental study as *Study One*. The following sections cover a summary of this study's motivation (Section 5.1), details on pre- and post-exclusion sample size (Section 5.2), a presentation of our planned statistical analyses and their results (Sections 5.3 and 5.4; as preregistered at **osf.io/pzja3**), and, finally, a report of certain unplanned analyses that we opted for only after having inspected the response data (Section 5.5).

5.1 Motivation

As already has been laid out in the prior chapters, we were motivated to conduct the present study by a desire to conceptually replicate the effect that is predicted by the polarity hypothesis proposed by van Tiel et al. (2019) and van Tiel and Pankratz (2021); see again Section 2.1.3. Beside this main hypothesis whose dependent measure of interest are response times during an SI-inducing verification task, we have also derived a secondary hypothesis (Section 2.1.4) that is concerned with how verification response choices themselves may be modulated by scalar polarity during such a task. We want to test both of these hypotheses using planned and rigorously hypothesis-driven analyses.

In addition, the scarcity of previous research on potential relationships between individual differences in WMC, fluid intelligence, or print exposure and the processing behaviour shown by human comprehenders when confronted with potential SIs (of varying polarity) motivated us to fill this gap by running this particular study which combines an SI-related task (our SPV implementation) with three cognitive tests that assess individual differences (OSpan, RPM, and ART).

5.2 Sample Size

Study One was run with a planned post-exclusion sample size of $N = 100$ participants. Choosing this particular sample size (which had been preregistered) was motivated informally by the intuition that at least a (bare-minimum) three-digit number of subjects would usually be necessary for meaningful assessments of individual differences.

Based on the stopping rule implied by this post-exclusion sample size, we eventually had to recruit overall 111 Prolific workers who made it through the entire experiment (pre-exclusion). Two of those subjects had to be manually excluded from analysis because they reportedly experienced a technical glitch that unintentionally had them skip parts of the SPV and OSpan tasks, thus rendering their response data incomplete and unusable. One further subject needed to be excluded manually as they admitted to cheating on the OSpan task by writing down letters in order to facilitate (or actually, bypass) memorisation. Eight more subjects were then automatically excluded from subsequent analyses based on the exclusion criteria defined back in Section 4.2; to be precise, three out of these eight subjects did not meet the SPV-based inclusion criterion, four of them did not meet the OSpan-based one, and there was one subject who failed to meet both of these aforementioned criteria. Hence, we do, indeed, end up with a post-exclusion data set that encompasses responses from exactly 100 individual subjects.

These 100 subjects display the following summary demographics: Their mean age at the time they took the study was 35 years, with a sample SD of 3 years. The minimum age was at 30 and the maximum age at 40 because sampling was constrained as described in Section 4.2. There were 48 men, 50 women, and 2 people of non-binary gender among the subjects (cf. biological sex: 46 males, 53 females, 1 undisclosed). While all subjects were native speakers of English, 13 of them had self-reportedly grown up as bilinguals, i.e., with more than one language.

5.3 Planned Hypothesis-Driven Analyses

This is an overview of the planned hypothesis-driven analyses conducted on the response data from the SPV task. Methodological aspects of these statistical analyses are discussed in Sections 5.3.1–5.3.5. Then, in Section 5.3.6, we report the results that these analyses yield once applying them to the collected data.

5.3.1 Dependent Measures

Two behavioural measures, i.e., response time and verification response, are considered as our dependent measures of interest in the statistical models described in Section 5.3.3. What follows here is a summary of their properties.

Response Time

Response time (RT) is measured as the time interval between the key press that makes a trial's picture appear and the subsequent conditional key press for verification of the trial as either 'good description' or 'bad description' (read: 'True' or 'False'). It is thus a continuous numerical variable. As is common practice with various kinds of psychological reaction-time data (see Whelan, 2008), we apply a log-transformation to the raw RT measures. Thus, to be very precise, our actual measure of interest is given by the natural logarithm of any raw RT that is given in milliseconds. We denote it as `log_RT` in the model specifications presented further below.

Verification Response

Whether a participant judges a given sentence–picture combination to represent a ‘good description’ or a ‘bad description’ (again, read: ‘True’ or ‘False’) is recorded as their verification response, thus giving rise to a binary variable. Here, as a dependent measure, it is coded as 1 = ‘True’ and 0 = ‘False’ so that it can be employed in a logistic mixed-effects model. It is abbreviated as `VR.bin` in the model specifications below.

5.3.2 Predictor Variables

Here follows a description of the three predictor variables that are used in the statistical models described below in Section 5.3.3.

Condition

Condition is a binary variable that distinguishes between *critical* (in Table 4.3: *t1*+ or *t1*−) and *control* (ibid.: *t2*+, ..., *t6*+ or *t2*−, ..., *t6*−) trials. But in order to keep statistical complexity in check, we decided to not actually include it as a factor in any of our hypothesis-driven models. Rather, we opt for an approach where we first fit a model only on the control trials and then use predictions from that model to residualise out variance from an eventually fitted model of interest on critical trials (for more on that, see Section 5.3.3).

Verification Response

This is the same variable that has already been summarised above in Section 5.3.1, but it is employed as a predictor variable (rather than a dependent measure) in those of our models that try to predict response time. In this function, it is sum-coded (rather than dummy-coded as 1 and 0) with regard to the data set of critical trials (with ‘True’ > 0 and ‘False’ < 0) and shall be abbreviated simply as `VR`.

Polarity

Polarity is another binary variable. It refers to whether the Horn scale underlying a trial is a *positively* or *negatively* polar one. Note that, here, this variable can vary within items as each item can appear in four different sentence variants, with two for each polarity level. This predictor variable is also sum-coded with regard to the data set of critical trials (with positive polarity > 0 and negative polarity < 0). Polarity is abbreviated as `POL` in the ensuing model specifications.

5.3.3 Statistical Models

The first analysis aims to replicate what is predicted by the polarity hypothesis, i.e., a B&N effect for positively polar scalars, but a reversed B&N effect for their negatively polar counterparts. In order to do that, two linear mixed-effects models—labelled Control Model and Model (i)—are constructed and implemented using the *lme4* package (Bates, Mächler, Bolker and Walker, 2015) for the R programming language (R Core Team, 2023). The models are fitted using the *bobyqa* optimiser. Their specifications are given below in *glmer* syntax. As specified here, a maximal random-effects structure would be used in each model.

But in case a model fails to converge, its random-effects structure is simplified, starting by removing by-item slopes for the interaction term, then (if necessary) by-subject slopes for the interaction term. If non-convergence still persists, by-item slopes for verification response, then (if still necessary) polarity are removed, and eventually analogous by-subject slopes in the same order.[15]

Control Model, on control trials:

- Gaussian distribution with identity link function
- `log_RT` $\sim$ `VR * POL + (VR * POL | subject) + (VR * POL | item)`

The Control Model's predicted log RTs are extracted and used in order to residualise out effects of verification response and polarity that are not specific to critical trials from Model (i) below. That is, the dependent measure for Model (i), fitted on critical trials, is the difference between the real log RT (on any critical trial) and the control-predicted log RT for the corresponding particular combination of subject, item, verification response, and polarity. This modified dependent measure is denoted as `res_log_RT`, here:

Model (i), on critical trials:

- Gaussian distribution with identity link function
- `res_log_RT` $\sim$ `VR * POL + (VR * POL | subject) + (VR * POL | item)`

For the second statistical analysis, which is designed to probe potential effects of scalar polarity *on* verification response choices, a logistic mixed-effects model—labelled Model (ii)—is constructed. Its implementation relies on the same software as the previous two models, and it is also fitted using the *bobyqa* optimiser.

Model (ii), on critical trials:

- Binomial distribution with logit link function
- `VR.bin` $\sim$ `POL + (POL | subject) + (POL | item)`

Again, this maximal random-effects structure, given as a starting point, is similarly simplified in case of non-convergence: First, by-item slopes for polarity would be removed, then by-subject slopes for polarity.[16]

In Section 5.3.4 right below, we discuss the predictions that are tested by fitting Models (i) and (ii). Two-sided statistical tests are carried out either using the Satterthwaite (1946) procedure for approximating degrees of freedom, in case of the *linear* Model (i), or using the Wald (1943) z-test, in case of the *logistic* Model (ii). Each statistical test is evaluated against a significance threshold of $p < .05$, where we correct for multiple $(= 2)$ comparisons using the Holm–Bonferroni method (Holm, 1979).

5.3.4 Hypotheses

The following predictions are made, which have already been laid out less formally back in Section 3.1:

(i) In Model (i), the interaction term `VR:POL` shows a significant effect of such directionality that pragmatically verified, positively polar trials as well as literally verified, negatively polar trials display additional delays in residual response time.

[15] Now, in hindsight, we realise that fixing this arbitrary deterministic order of removing random-effect slopes one after the other (and with them, any random-correlation terms) until convergence happens is not exactly an ideal approach. Instead, in a future study, we would probably preregister an approach where the selection of a final random-effects structure is guided by considering which of the random-slope parameters in a too complex model accounts for the least amount of variance and can, thus, most reliably be dropped; additionally, we would also try constraining random-correlation terms to zero.

[16] See footnote 15.

(ii) In Model (ii), the main-effect term `POL` shows a significant effect of such directionality that positively polar trials display a stronger bias towards being verified literally than negatively polar trials.

It should be noted that, although we have these clear expectations about the sign of either effect in (i) and (ii), we nonetheless choose to perform two-sided and, thus, rather conservative statistical tests.

5.3.5 Power Simulations

In order to acquire estimates of how likely we would be, given our sample size, to find either of our two predicted effects if it was, indeed, real, we carried out a prospective power analysis for each of the effects using simulations. It should be noted that these two power analyses did not actually influence our decision on sample size for Study One, here, which had already been fixed at the time ($N = 100$). Rather, we conducted them just as an additional exercise and also in the hope that they may provide further insights in light of which our eventual results would be more sensibly interpretable.

The power analyses were implemented using the *simr* package (Green and MacLeod, 2016) for the R programming language (R Core Team, 2023). For reproducibility, our corresponding R script can be found here: **osf.io/yc8tm**.

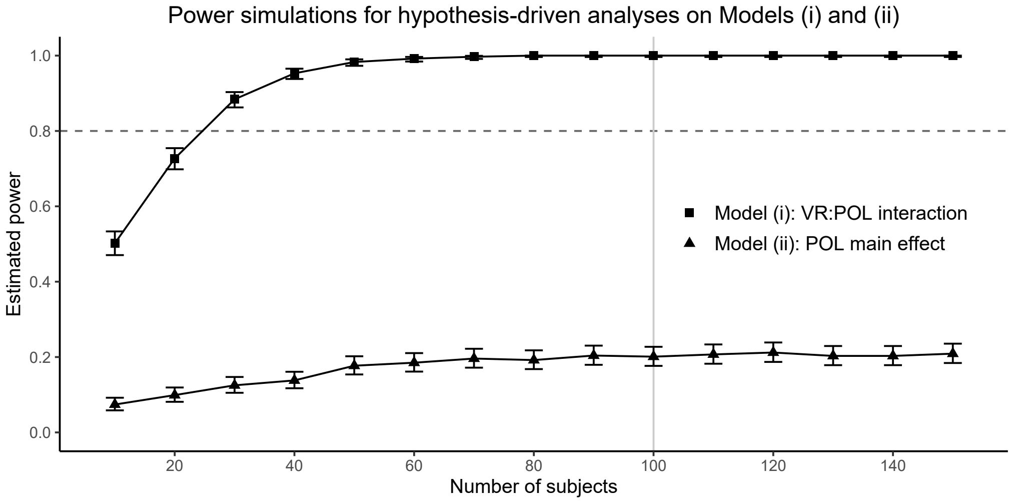


Figure 5.1: Estimated power for detecting `VR:POL` interaction effect on residual log RT in linear mixed-effects Model (i) is represented by the plot consisting of interconnected, quadratically shaped dots. Further, estimated power for detecting `POL` main effect on verification response in logistic mixed-effects Model (ii) is represented by the plot consisting of interconnected, triangularly shaped dots. Different sample sizes from 10 up to 150 are considered. Underlying effect-size estimates are drawn from response data from overall only 20 pilot subjects that was collected on 16 February 2023. 1,000 simulations were run for each probed sample size. Error bars represent 95 % CIs. The horizontal, dark grey, dashed line marks the conventional power threshold of 0.8. The vertical, light grey, continuous line highlights the sample size that was chosen for Study One ($N = 100$) independently of these simulation-based power estimates.

We considered the 15 possible sample sizes $N \in \{10, 20, \ldots, 150\}$. The parameter `nsim` (number of simulations) was set to 1,000; that is, for every probed sample size, we fitted Models (i) and (ii) each on 1,000 independently simulated data sets. The estimate of statistical power is then simply the ratio of simulations where the particular effect of interest is found to be significant; for example, if significance is found in 600 out of 1,000 simulations, the resulting power estimate is 0.6.

Importantly, the two power analyses were based on the assumption that the effect-size estimates that can be found in our $N = 20$ pilot data from 16 February (see Section 4.4 and for corresponding data files: **osf.io/mqpk2**)[17] are reasonable approximations of any real effects 'out there in the world' that we want to find using our Models (i) and (ii). However, perhaps by chance due to our small pilot sample size (i.e., Type M error; see Gelman and Carlin, 2014), we find a highly optimistic estimate of the effect size of interest in Model (i) in the given pilot data ($\hat{\beta}_{\texttt{VR:POL}} = -0.37$, on the log-milliseconds scale).[18] Hence, curiously, our present power analysis suggests that already with a sample size of $N \geq 60$ we would attain ≥ 0.99 statistical power for detecting our desired `VR:POL` interaction effect on residual log RT in Model (i), as illustrated by Figure 5.1. With respect to Model (ii) and its binary dependent measure (verification response), the effect-size estimate for our expected `POL` main effect is quite small in the pilot data (i.e., $\hat{\beta}_{\texttt{POL}} = -0.17$, on the log-odds scale).[19] Consequently, the present power analysis which is based on it ends up being hugely pessimistic about finding such an effect with any of the probed sample sizes up to $N = 150$. As Figure 5.1 shows, the corresponding power curve stagnates around 0.2. Typically, statistical power of at least 0.8 is considered desirable.

In conclusion, even though they are interesting to look at, these power analyses should be taken with a grain of salt as the effect-size estimates they are based upon might be spurious given that they were drawn from small-sample pilot data.

5.3.6 Results

Let us now turn to the results of our hypothesis-driven analyses, carried out on Study One's SPV response data from 6 and 7 March 2023. They can be reproduced by running the script stored in **osf.io/65chg** on the data in **osf.io/yd8mc** until code line 684.

Both of our predictions (Section 5.3.4) fail to be validated by the data: The hypothesised `VR:POL` interaction effect on residual log RT in Model (i) is not significant ($\hat{\beta}_{\texttt{VR:POL}} = -0.10$, $SE = 0.07$, $t = -1.53$, $p_{\text{raw}} = .127$, $p_{\text{Holm–Bonferroni}} = .254 \geq .05$). Similarly, the hypothesised `POL` main effect on verification response in Model (ii) is not significant either ($\hat{\beta}_{\texttt{POL}} = -0.46$, $SE = 0.87$, $z = -0.53$, $p_{\text{raw}} = .595$, $p_{\text{Holm–Bonferroni}} = .595 \geq .05$). Ergo, in Study One, we neither obtain support for the polarity hypothesis (Section 2.1.3) nor for our secondary, polarity-based hypothesis (Section 2.1.4).

An overview of all fixed-effect parameter estimates produced by the fitted Models (i) and (ii) can be found in Tables 5.1 and 5.2, respectively. Based on these, we can make some exploratory observations:

In Model (i), the intercept estimate is at $\hat{\alpha} = 0.37$ ($SE = 0.05$). Recall that this represents the dependent measure residual log RT as defined in Section 5.3.3: A residual log RT of 0.37 log ms implies a relative 0.37-log-millisecond slowdown on a critical trial as compared

[17] Why do we consider the data from our third pilot study (16 February) in particular, but not from the two pilot studies that preceded it (28 January, 13 February)? Because only in the third pilot study the experimental design of the SPV task is exactly identical to the one later also used in Study One.

[18] Notably, the effect in the pilot data even has the expected sign (i.e., $\hat{\beta}_{\texttt{VR:POL}} < 0$, given the contrast-coding scheme defined back in Section 5.3.2).

[19] That is, the model fitted on the pilot data suggests that positively polar SPV trials are $e^{-0.17} = 0.84$ times as likely to receive a *literal* response as negatively polar trials. Or, put the other way around, positively polar SPV trials are $e^{0.17} = 1.18$ times as likely to receive a *pragmatic* response as negatively polar trials.

Table 5.1: Fixed-effect estimates from linear mixed-effects Model (i), fitted to residual log RT, based on SPV response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). Corresponding random-effect estimates from the same model are summarised further below in Appendix B, in Table B.1. Moreover, this table has a twin further below in Section 6.3.2 where analogous results from a follow-up study with larger sample size are discussed: Table 6.2.

	$\hat{\alpha}$	*SE*	t	p
Intercept	0.37	0.05	7.38	.001
	$\hat{\beta}$	*SE*	t	p
Verification response (`VR`)	−0.03	0.04	0.82	.412
Polarity (`POL`)	−0.05	0.05	−1.10	.330
Interaction of `VR:POL`	−0.10	0.07	−1.53	.127

Table 5.2: Fixed-effect estimates from logistic mixed-effects Model (ii), fitted to (critical-trial) verification response, based on SPV response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). Corresponding random-effect estimates from the same model are summarised further below in Appendix B, in Table B.2.

	$\hat{\alpha}$	*SE*	z	p
Intercept	0.52	0.43	1.22	.224
	$\hat{\beta}$	*SE*	z	p
Polarity (`POL`)	−0.46	0.87	−0.53	.595

to a similar control trial. And since we have sum-coded our predictors with respect to the critical-trials data set, we can interpret the intercept estimate $\hat{\alpha}$ directly as the predicted mean value for residual log RT. A sanity-check comparison against the trivially calculated grand mean value of residual log RT (which is ≈ 0.37 log ms as well) confirms this. For easier interpretability, consider that a value of 0.37 log ms does roughly correspond to 792 ms around the grand mean RT (2,128 ms) found in our complete (critical + control) experimental data set. The main-effect term of verification response (`VR`) displays a slope estimate of only $\hat{\beta}_{\texttt{VR}} = -0.03$ ($SE = 0.04$), i.e., it seems to be barely accounting for any variation in residual log RT at all. Something similar is the case for the main-effect term of polarity (`POL`), with only $\hat{\beta}_{\texttt{POL}} = -0.05$ ($SE = 0.05$).

In Model (ii), the only fixed-effect term apart from our `POL` main effect of interest is the intercept. It is at $\hat{\alpha} = 0.52$ ($SE = 0.43$) which, as you may recall, represents the log odds of giving a literal response as opposed to a pragmatic one. If the uncertainty associated with this estimate were not so huge, then it could, just in principle, be interpreted as indicating that literal (‘True’) responses to critical trials are generally $e^{0.52} = 1.68$ times as likely to occur as pragmatic (‘False’) responses.

Regarding random-effects structure, the non-singularly converging Model (i) variant that we have eventually chosen, here, based on our predefined deterministic selection criterion (see Section 5.3.3) features by-subject and by-item random intercepts as well as by-subject random slopes for verification response and for polarity (but not their interaction) and by-item slopes only for polarity. All accordingly possible intercept–slope or slope–slope correlation terms are also included as random parameters. However, in the employed Control Model variant—which was fitted on control trials in order to yield predicted log RTs so as to obtain residual log RTs within the data set of critical trials—only by-subject and by-item random intercepts, but no random slopes whatsoever had been included.

The random-effects structure of our employed, converging Model (ii) variant is maximal, i.e., it includes by-subject and by-item random intercepts as well as by-subject and by-item random slopes for the single predictor polarity. Note that by-subject and by-item intercept–slope correlation terms are included as random parameters as well.

A complementary overview of random-effect parameter estimates from the fitted Models (i) and (ii) is provided in Appendix B, in Tables B.1 and B.2, respectively.

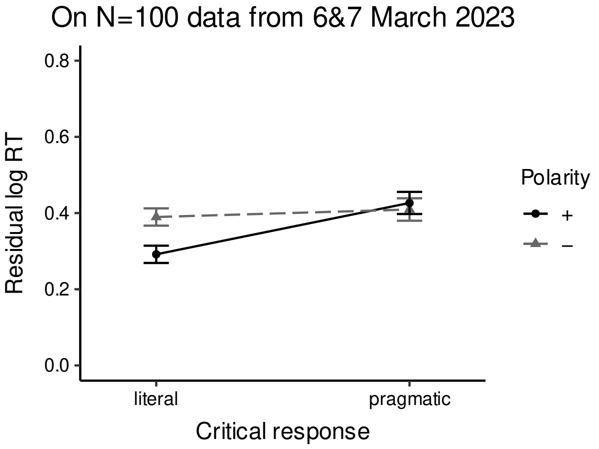


Figure 5.2: Mean residual log RTs grouped by (critical-trial) verification response and by polarity. The displayed error bars represent within-subject-and-item standard errors (computed following Morey et al., 2008). The underlying SPV response data was collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). One is invited to compare this plot to a sister plot in Appendix B (Figure B.1) where log RTs rather than residual log RTs are displayed based on the same data. Moreover, there is a twin plot further below in Section 6.3.2 where analogous results from a follow-up study with larger sample size are presented: Figure 6.3.

To add some visual intuition to our obtained main results, please first consider the plot in Figure 5.2. It displays mean values of residual log RT by (critical-trial) verification response and by polarity. As can be noticed, with positively polar trials, there is a slight upwards slope going from literal (0.29 log ms) to pragmatic responses (0.43 log ms). By contrast, with negatively polar trials, residual log RTs are almost the same, on average, for both literal (0.39 log ms) and pragmatic responses (0.41 log ms). The shape and direction of this pattern are, in principle, consistent with the theoretical predictions by some version of the polarity hypothesis (here: present B&N effect for positive polarity, but *absent* B&N effect for negative polarity). However, as we saw in Table 5.1, the effect magnitude ($\hat{\beta}_{\texttt{VR:POL}} = -0.10$) associated with this interaction of interest is simply too small, in combination with the substantial uncertainty of the estimate ($SE = 0.07$), to consider it evidence ($p \geq .05$).

Now, compare Figure 5.2 with Figure 5.3 which contains the same kind of information as the former, but where that information has also been grouped by item, i.e., into four item-specific subplots. Quite strikingly, effect patterns appear to be hugely different across the examined items: E.g., for the Quantity and Space items, the plot resembles a '>' shape, with negative polarity showing the upper, but downward literal-to-pragmatic

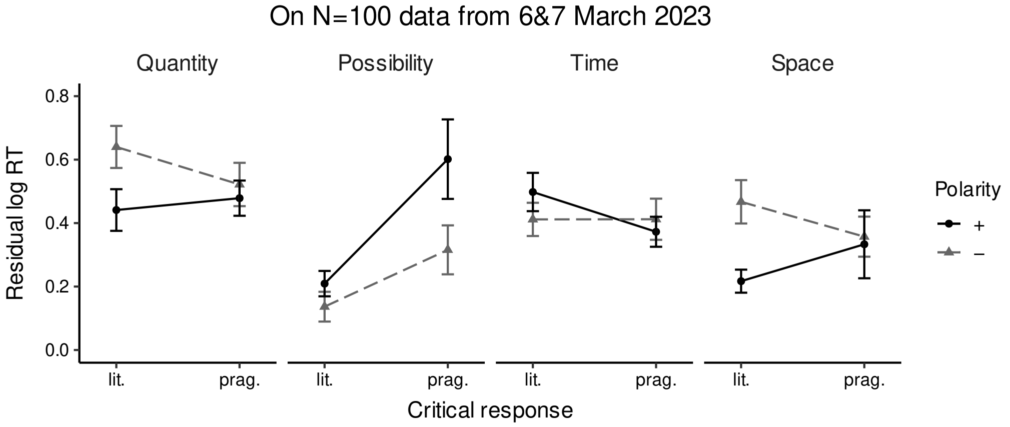


Figure 5.3: For each item separately: Mean residual log RTs grouped by (critical-trial) verification response and by polarity. The displayed error bars represent within-subject standard errors (computed following Morey et al., 2008). The underlying SPV response data was collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). One is invited to compare this plot to a sister plot in Appendix B (Figure B.2) where log RTs rather than residual log RTs are displayed based on the same data. Moreover, there is a twin plot further below in Section 6.3.2 where analogous results from a follow-up study with larger sample size are presented: Figure 6.4.

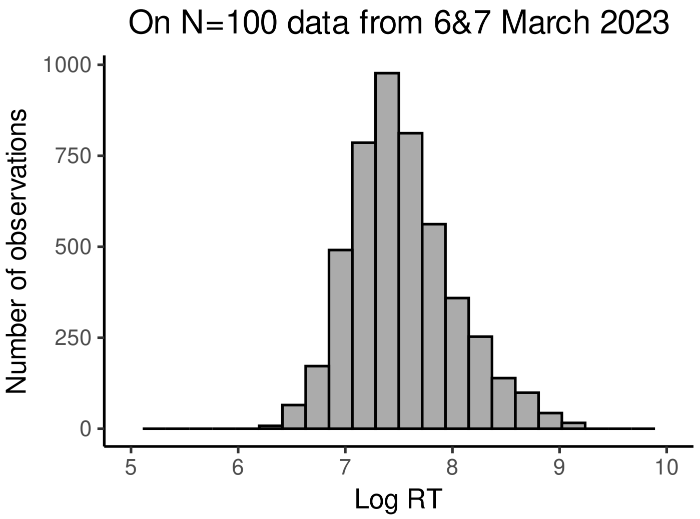


Figure 5.4: A histogram that shows the distribution of log RTs in the full (critical + control) data set of SPV response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). Here, log RT stands for the natural logarithm of a raw response time given in milliseconds. There is a twin plot further below in Section 6.3.2 where analogous results from a follow-up study with larger sample size are presented: Figure 6.6.

slope, thus vaguely resembling the directionality predicted by the polarity hypothesis. But for the Possibility item, we see a slightly tilted ‘<’ shape, with residual log RT for either polarity being larger for pragmatic responses than for literal ones—while this difference is more extreme for positive polarity.

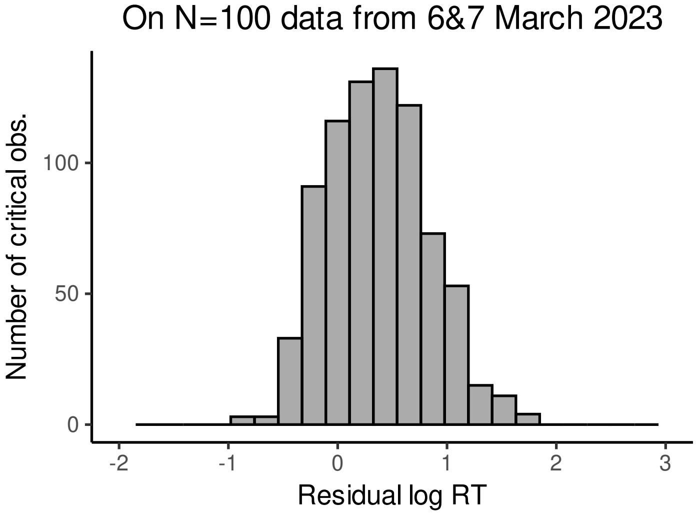


Figure 5.5: A histogram that shows the distribution of residual log RTs in the data set of critical-trial SPV response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). Residual log RT is calculated by subtracting a Control-Model-predicted value for log RT from the observed log RT of a critical trial (see Section 5.3.3). In this context, log RT stands for the natural logarithm of a raw response time given in milliseconds. There is a twin plot further below in Section 6.3.2 where analogous results from a follow-up study with larger sample size are presented: Figure 6.7.

The pattern is not so clear in case of the Time item due to somewhat overlapping error bars across polarity. The 16 plotted group means of residual log RT range from 0.14 log ms (Possibility item, negative polarity, literal response) to 0.64 log ms (Quantity item, negative polarity, literal response). Overall, this observed variability seems very relevant to consider when interpreting our main result: Although our Model (i) of residual log RTs can account for by-item variation to some extent thanks to including by-item random effects, this flexibility has obvious limits when it comes to trying to infer some salient general pattern that describes the overall data well despite (moderate) between-item differences, especially in a design like ours with very few (= 4) items. Moreover, the presently observed between-item differences might themselves be interesting to further investigate. They may arise due to any of multiple reasons, e.g., properties of the underlying Horn scales, of the particular sentences, or of the particular pictures involved. Of course, if they turned out to be grounded in genuine differences between Horn scales, that would be theoretically most interesting to pursue further in future work.

Generally, it should be conceded that interpreting residual log RTs is a bit less straightforward than directly interpreting (simple) log RTs for either experimental condition (critical/control). Therefore, in Appendix B, one can also find alternative variants of the two figures just discussed, where mean log RTs rather than mean residual log RTs are plotted instead, but also grouped by condition: Figures B.1 and B.2.

Figures 5.4 and 5.5 show histograms of the distributions of log RTs (critical + control trials) and residual log RTs (critical trials), respectively. Visually, both distributions appear to be approximately Gaussian (although log RTs have a slight positive skew), hence validating the underlying normality assumption of the Control Model and of Model (i) in retrospect.

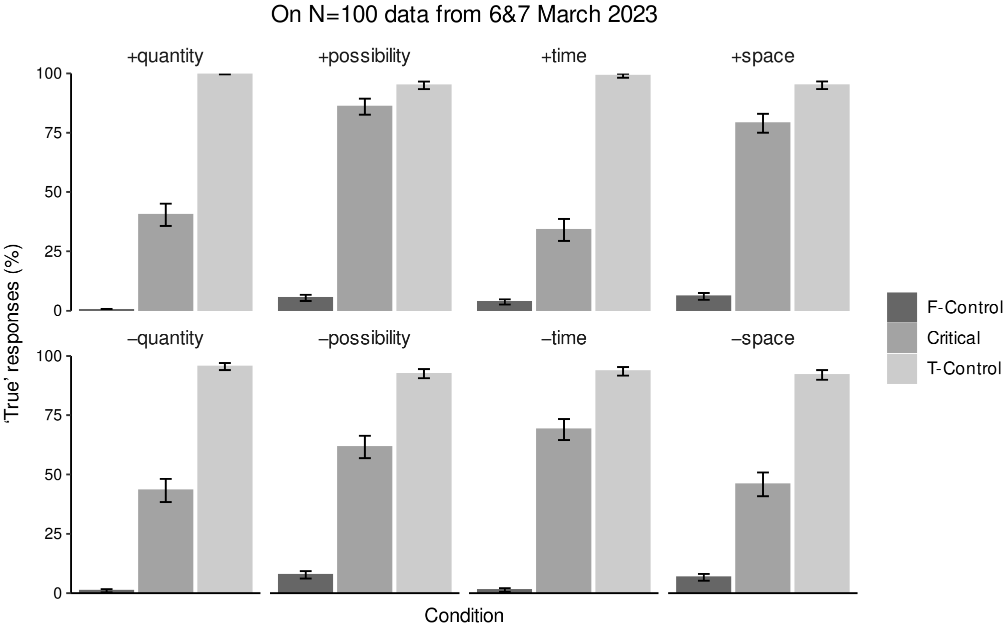


Figure 5.6: Bar plots showing ratios of 'True' responses in the SPV data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion), grouped by scale (+quantity, +possibility, ..., −space) and by three different condition set-ups: control trials where the expected response is 'False' (F-Control), critical trials (Critical), and control trials with a 'True' expected response (T-Control). The displayed error bars represent within-subject standard errors (computed following Morey et al., 2008). Note that there is a twin figure further below in Section 6.3.2 where analogous results from a follow-up study with larger sample size are presented: Figure 6.5.

Having visually inspected the response-time data, it would now also be useful to take a look at the 'True' vs. 'False' verification responses themselves and how their ratios differ by condition, polarity, and item. Figure 5.6 summarises this: For each particular scale (+quantity, +possibility, ..., −space), it shows a bar plot where each of three bars represents a different condition set-up: control trials where the expected response is 'False' (F-Control), critical trials (Critical), and control trials with a 'True' expected response (T-Control). The height of each bar represents the ratio of given 'True' responses. It seems interesting that the ratios for the critical condition vary quite a lot across items (e.g., with +quantity and −quantity having rather low ratios, 40 % and 43 %, but +possibility and −possibility having rather high ratios, 86 % and 62 %), but also within-item between-polarity (e.g., with +time having a substantially lower ratio, 34 %, than −time, 69 %, but +space having a substantially higher ratio, 79 %, than −space, 46 %). Especially the latter kind of variation, i.e., within-item between-polarity, illustrates very well why given the present results we cannot expect to find a significant main effect of polarity on verification response in Model (ii). The formerly mentioned kind of variation, i.e., simply across items, may motivate similar considerations that have already been discussed above regarding the substantial between-item variability in response-time patterns: Perhaps there are theoretical reasons why our examined scales (or, more pessimistically, our particular sentence–picture materials) provoke differential processing behaviour in comprehenders. An example of such a speculation could be that (modal) verbal expressions as in ⟨might be, is definitely⟩ display a stronger bias towards being interpreted literally due to their particular syntactic role, compared to quantifier pronouns as in ⟨some, all⟩.

5.4 Planned Exploratory Analyses

This section provides an overview of the planned, but exploratory (i.e., not hypothesis-driven) analyses that have been conducted as part of Study One. They seek to detect potentially relevant effects of individual differences (IDs) in WMC, fluid intelligence, or print exposure on (polarity-contrastive) SI processing behaviour.

5.4.1 Individual-Differences Variables

What follows is a summary of the properties of three variables resulting from measures that each operationalise one of the psychological constructs WMC, fluid intelligence, and print exposure, as introduced back in Section 2.2.1.

Working Memory Capacity

Working memory capacity (or WMC, as already abbreviated earlier) is a continuous numerical variable whose value is the score that a participant has achieved in the OSpan task. Calculated, here, using the partial credit unit (PCU) scoring procedure, it is the mean proportion of letters in a set (across all sets) that a participant has recalled correctly. It can therefore range between 0.0 and 1.0, but, for the present purposes, the values are standardised so that the sample mean becomes 0 and the sample SD becomes 1. In model specifications, here, it is abbreviated as `ID_WMC`.

Fluid Intelligence

Fluid intelligence is a discrete numerical variable whose value is the score that a participant has achieved in the RPM task. That score is simply the number of missing pieces (here, among ten trials) selected correctly; hence it can range from 0 to 10. For further analysis, here, its values are standardised so that the sample mean becomes 0 and the sample SD becomes 1. In model specifications, it is denoted as `ID_GF`.

Print Exposure

Print exposure is a discrete numerical variable whose value is the score that a participant has achieved in the author recognition test. That score is acquired by subtracting the total number of fake author names (falsely) classified as real ones from the total number of real author names that have been correctly classified. As participants with negative scores are excluded from the analysis (see Section 4.2), this variable can range between 0 and 65, but the values are standardised so that the sample mean becomes 0 and the sample SD becomes 1, for downstream analysis. We denote it as `ID_PE` in model specifications given below.

5.4.2 Model Selection

Recall Models (i) and (ii), i.e., those on critical trials, from Section 5.3.3. Separately for each of the three ID measures of interest, each of these two models is now extended by adding fixed effects for the particular measure as well as for all possible interaction terms it can be part of. In consequence, one ends up with the following six new models, all fitted on the critical-trial data set:

1. **Model (i-WMC-full),** Gaussian distribution with identity link function:
   - `res_log_RT`
     `∼ VR * POL * ID_WMC + (VR * POL | subject) + (VR * POL | item)`
2. **Model (i-GF-full),** Gaussian distribution with identity link function:
   - `res_log_RT`
     `∼ VR * POL * ID_GF + (VR * POL | subject) + (VR * POL | item)`
3. **Model (i-PE-full),** Gaussian distribution with identity link function:
   - `res_log_RT`
     `∼ VR * POL * ID_PE + (VR * POL | subject) + (VR * POL | item)`
4. **Model (ii-WMC-full),** binomial distribution with logit link function:
   - `VR.bin ∼ POL * ID_WMC + (POL | subject) + (POL | item)`
5. **Model (ii-GF-full),** binomial distribution with logit link function:
   - `VR.bin ∼ POL * ID_GF + (POL | subject) + (POL | item)`
6. **Model (ii-PE-full),** binomial distribution with logit link function:
   - `VR.bin ∼ POL * ID_PE + (POL | subject) + (POL | item)`

These six models may be referred to as the *theoretical full* models. Any theoretical full model's random-effects structure is now simplified step-by-step by removing random slopes, as far as necessary, in the same order as described earlier in Section 5.3.3, until convergence is ensured. For cases where a model would still not converge, though, even after removing all random slopes, we had planned (preregistration: **osf.io/pzja3**) to then also try dropping some of its interaction-term *fixed* effects, but this scenario actually never came up given our eventually collected data. The resulting converging models, each with its possibly somewhat simplified parameter structure, may be referred to as the *actual full* models hereinafter.

Then, a process of backwards model selection is performed: Using likelihood ratio tests (at a significance threshold of uncorrected $p < .05$), it is checked whether the removal of any ID predictor (main-effect or interaction term) from the fixed effects significantly worsens model fit. If this is the case, the predictor is kept in the model. If, however, model fit stays mainly unaffected by the removal of a predictor, then that predictor is dropped. Such probing of the relevance of predictors is done in the following order (ignoring predictors that are absent in the particular actual full model anyway): `VR:POL:ID_⟨…⟩` → `VR:ID_⟨…⟩` → `POL:ID_⟨…⟩` → `ID_⟨…⟩`.

As a result, there have now been selected six modified models—one may call them the *reduced* models—that each contain between zero and four ID predictors (main-effect or interaction terms, all involving the same particular ID measure) among their fixed effects. We label these as Models (i-WMC-reduced), (i-GF-reduced), …, (ii-PE-reduced). The presence or absence of any such predictor in any of these reduced models could now allow for exploratory conclusions regarding the question if and how particular ID measures modulate the processing of variously polar SIs.

5.4.3 Results

In this section, we report the results of the planned exploratory analyses just described. Recall that they rely on the response data from four different tasks, i.e., SPV and three

ID tests, administered on 6 and 7 March 2023 to 100 Prolific participants (post-exclusion). These results can be reproduced by running the script stored in **osf.io/65chg** on the data in **osf.io/yd8mc** up until code line 998.

Strikingly, among the overall 18 statistical comparisons that have been performed here, probing, for each of the considered ID-based fixed effects, if it improves model fit in variants of Models (i) and (ii), only two end up with a positive result: The main effect of fluid intelligence (RPM scores) is found to modulate residual log RT ($\chi^2 = 8.06$, $p = .005$) insofar as individuals with higher fluid intelligence tend to display larger residual log RTs ($\hat{\beta}_{\texttt{ID\_GF}} = 0.06$); in other words, a trial's being a critical rather than a control one makes such individuals slow down even more than their less fluidly intelligent peers. Secondly, including a three-way interaction between verification response, polarity, and print exposure (ART scores) also improves model fit on residual log RT ($\chi^2 = 4.23$, $p = .040$), in such a direction ($\hat{\beta}_{\texttt{VR:POL:ID\_PE}} = 0.13$) that, seemingly, something approximating the two-way, `VR:POL` effect pattern predicted by the polarity hypothesis is *more* salient *the less* an individual has been exposed to print in the past. Plots of model predictions that highlight the directional patterns of these two potential effects can be inspected in Figures 5.7 and 5.8. An overview of all the present model-selection results, including BIC goodness-of-fit values (Schwarz, 1978) for actual full, for reduced, and for *original* (i.e., as fitted in Section 5.3) variants of Models (i) and (ii), is given in Table 5.3.

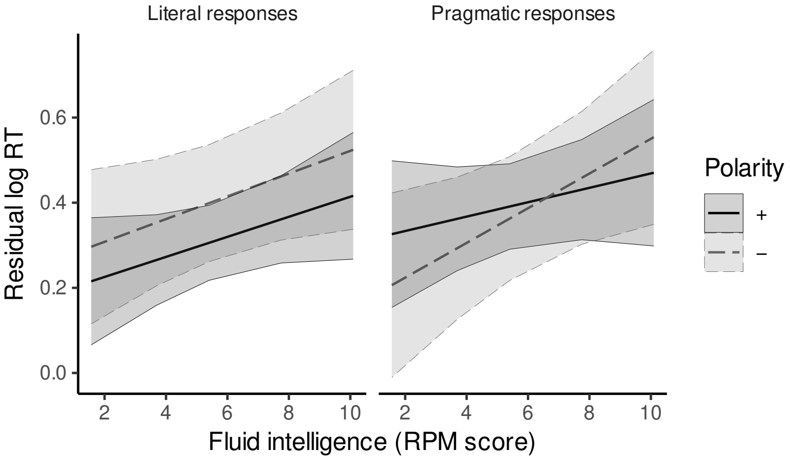


Figure 5.7: Predictions of the fitted Model (i-GF-full), illustrating the main effect of fluid intelligence, as operationalised through RPM scores, on residual log RT. Dark grey (positive polarity) or light grey (negative polarity) shaded areas are error ribbons representing 95 % CIs.

Of course, this very necessary disclaimer has to be put forward: Because this is an exploratory analysis in the sense that its multiple (= 18) comparisons are not corrected for at all,[20] caution is advised before eagerly drawing interpretative conclusions about real-world cognitive phenomena just based on the two present 'significant' results concerning fluid intelligence and print exposure. Nevertheless, what these results can actually offer us, is a vague theoretical direction regarding what might and might not be worth looking into whenever designing a future, confirmatory study about ID effects in (polarity-contrastive) SI processing. Our best available guess, then, beyond what prior theory and

[20] In fact, under the scenario that there was no real ID effect whatsoever among all kinds that we have probed, one would still be more likely to end up with at least some (false-positive) significant result, i.e., $1 - (1 - 0.05)^{18} \approx 60.3\,\%$ likely, than to actually only observe null results, here.

prior experimental work has already been able to suggest to us (Section 2.2), would be to set up a confirmatory study on IDs, SIs, and polarity, where (only) response time is considered as a dependent measure and (only) potential ID effects grounded in fluid intelligence and print exposure are examined, with explicit and, ideally, theoretically justified hypotheses about each. Especially in the simpler case of the observed main effect of fluid intelligence, plausible theoretical explanations can be thought of: For instance, it may be the case that fluidly more intelligent people are more sensitive to pragmatic ambiguities such as those caused by SIs (or more aware of and conflicted by the response dilemma they pose) and therefore tend to display additional temporal delays in their responses to them.

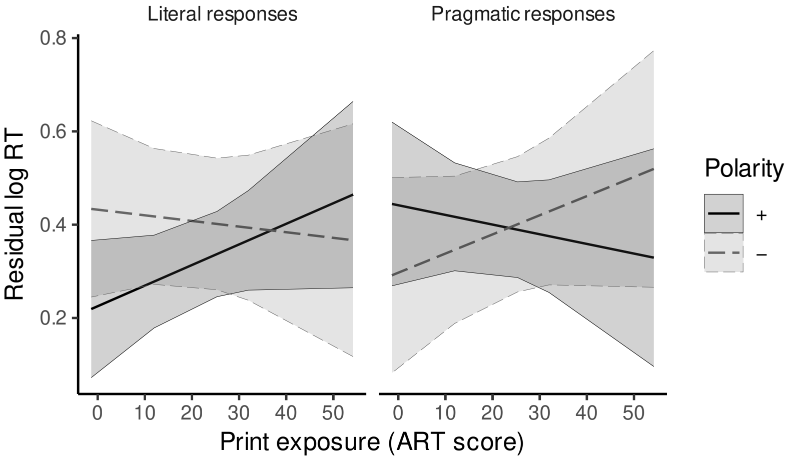


Figure 5.8: Predictions of the fitted Model (i-PE-full), illustrating the three-way interaction between verification response, polarity, and print exposure, as operationalised through ART scores, on residual log RT. Dark grey (positive polarity) or light grey (negative polarity) shaded areas are error ribbons representing 95 % CIs.

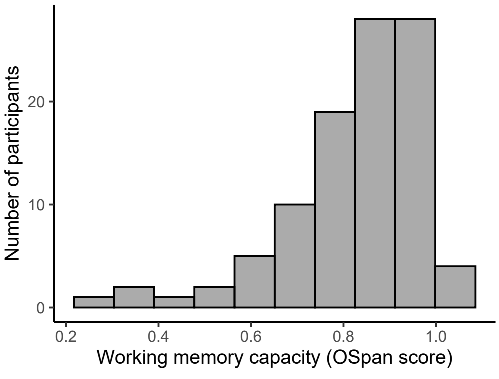


Figure 5.9: A histogram showing the distribution of OSpan performance scores, operationalising working memory capacity, as found in the response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion).

Table 5.3: A summary of the results yielded by the process of backwards model selection specified in Section 5.4.2, for each combination of model type and ID measure. These results stem from the response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). Reported χ^2 statistics (and their associated p values) for the two ID-based fixed effects that were retained after respective model reductions are derived from the performed likelihood ratio tests. In addition, absolute BIC goodness-of-fit values for the original Models (i) and (ii), fitted without including IDs, are given in the rightmost column; relative to these, increases or decreases in BIC for the corresponding reduced or full model variants can be found in the columns second and third from the right, respectively. Crucially, note that *lower* BIC values represent *better* goodness of fit, and that BIC penalises adding complexity to a model if that does not lead to a substantially larger amount of variance being accounted for.

Model	**ID Measure**	**Retained ID Fixed Effects**	**BIC**		
			Full	Reduced	Original
(i)	WMC	—	$+23$	± 0	1,063
	Fluid intelligence	`ID_GF` ↪ $\hat{\beta} = 0.06$, $\chi^2 = 8.06$, $p = .005$	$+18$	-1	1,063
	Print exposure	`VR:POL:ID_PE` ↪ $\hat{\beta} = 0.13$, $\chi^2 = 4.23$, $p = .040$	$+21$	$+2$	1,063
(ii)	WMC	—	$+11$	± 0	958
	Fluid intelligence	—	$+10$	± 0	958
	Print exposure	—	$+11$	± 0	958

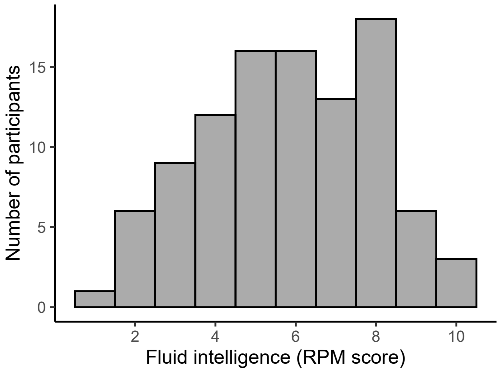


Figure 5.10: A histogram showing the distribution of RPM performance scores, operationalising fluid intelligence, as found in the response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion).

To complement our presentation of the results obtained from the model selections, we now turn to some descriptive statistical properties displayed by the collected ID-measure data: The performance scores from the three administered ID tests show very low intercorrelations among each other (Pearson's coefficients: $\hat{\rho}_{\texttt{ID\_WMC},\texttt{ID\_GF}} = 0.22$, $\hat{\rho}_{\texttt{ID\_WMC},\texttt{ID\_PE}} = -0.04$, $\hat{\rho}_{\texttt{ID\_GF},\texttt{ID\_PE}} = 0.13$). This may be seen as 'good news' in the sense that they do appear to, indeed, measure three separate underlying cognitive constructs of interest. In Figures 5.9, 5.10, and 5.11, histograms of the distributions of each measure's scores have been plotted. Interestingly, none of the three distributions seem to resemble a Gaussian one very closely: The OSpan scores are distributed along a strong negative skew, with the mean value of 0.82 (standardised: 0) being considerably closer to the maximum value of 1 (standardised: 1.15) than to the minimum value of 0.22 (standardised: -3.88).

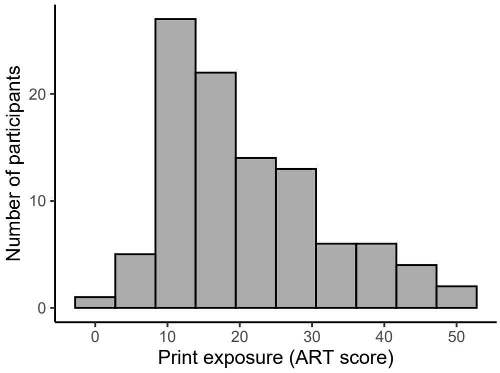


Figure 5.11: A histogram showing the distribution of ART performance scores, operationalising print exposure, as found in the response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion).

The large number of high-performance scores may hint at ceiling effects, similarly as observed by Mayn and Demberg (2022). This is despite the fact that, in the present study, an alternative Lingoturk implementation featuring larger set sizes is employed (see Section 4.1.2), designed afterwards by the main author Mayn to avoid such ceiling effects. In any case, the ceiling effects do seem to have been mitigated somewhat by this adapted implementation, employed here, as the mean score is already somewhat lower than the more extreme ones that had been found by Mayn and Demberg (i.e., 0.92 and 0.93) in their $N = 65$ and $N = 68$ experiments. The RPM score distribution exhibits a different peculiar pattern as it has two salient local maxima: one at a score of 5 or 6 (standardised: -0.39 or 0.08), with 16 participants each, that actually seems like the 'natural' maximum given the overall shape of the curve, assuming it should be Gaussian, but also another, actually the global one, at a score of 8 (standardised: 1.02), with 18 participants. We suspect that this second, even steeper peak which looks somewhat anomalous given the overall shape of the distribution might be explained by the fact that a salient minority of the examined participants has had previous experience with the RPM task—which is a quite common and anecdotally well-recognised way of administering 'IQ tests'. That is, the subset of those RPM-familiar participants might form a separate distribution with a mean score around 8, while the remaining, RPM-agnostic participants form the majority distribution centred around a score of 5 or 6. What supports this speculation is that on the task-specific, post-completion question slide for the RPM task (described in Section 4.3), 18 % of participants reported to have encountered some variant of this task before, a remarkably higher ratio than for any of the two other ID tests. And indeed, fitting a simple linear model to predict (raw, i.e., 0, 1, 2, ..., 10) RPM scores from the (treatment-coded) factor whether a subject did (= 1) or did not (= 0) report to have seen the task in some form in the past does indicate a notable relationship of the kind that we suspect ($b = 1.36$, $t = 2.51$, $p = .014$), with the mode of scores from the subsample of RPM-agnostic subjects being 5, but the mode of scores from RPM-familiar subjects being 8. This observation highlights the challenge posed by too popular operationalisations of cognitive constructs (like the RPM task for fluid intelligence) due to their high likelihood of being administered to non-naïve participants. It also calls for follow-up work in the present research context to consider alternative, less often administered ways of assessing someone's fluid intelligence. Finally, the distribution of scores from the ART task shows a noticeable positive skew, with a long tail to the right after it already peaks around scores between 10 and 15, and where the mean value of 20.86 (standardised: 0) is rather close to the minimum value of 0 (standardised: -1.88), but is a bit further away from the maximum value of 50 (standardised: 2.62).

5.5 Unplanned Exploratory Analyses

Now, we move on to reporting two further kinds of conducted analyses that were motivated and implemented only after running the experiment of Study One and inspecting its collected response data from 6 and 7 March 2023. In both cases, the purpose is to offer some additional, subtle insights that are not yet provided by the analyses we had planned in advance and have just reported in the earlier sections.

5.5.1 Examining the Literal Responder Group

When it comes to assessing individual-level differences, it is not only useful to probe if and how performances in some cognitive task correlate with some domain-specific measure of interest. Rather, it can also already be insightful to check how subjects are distributed based on average values of that dependent measure of interest itself. Hence, in our case, we would like to get an impression of whether there are any distinguishable chunks or clusters of responder types, based on their individual degree of inclination either towards producing literal or rather towards producing pragmatic responses in the SPV task.

Such an impression is provided visually by the histogram in Figure 5.12: As it shows, when comparing participant-specific ratios of pragmatic responses, there does not seem to be any clear-cut latent grouping into different kinds of responder groups. That is, the resulting distribution of such ratios has a quite steady shape rather than, say, displaying two separate peaks at its extreme (i.e., most literal and most pragmatic) regions. However, what we can acknowledge is that, given our data, pragmatic 'extremists' seem to be a much scarcer breed than literal 'extremists': We only find one subject who responded pragmatically throughout the entire task (i.e., 100 % pragmatic ratio, due to 8 pragmatic responses in all 8 critical trials). By contrast, there were a total of 11 subjects who always gave a literal response (i.e., 0 % pragmatic ratio).

Although, as we just saw, there is no visually obvious way of clustering our participants into responder groups based on their ratios of pragmatic responses, we may also arbitrarily define some cut-off values to obtain such a grouping using simple conceptual assumptions. Let us work with the following: We shall consider subjects with a pragmatic ratio smaller than $33.\overline{3}$ % as *literal-leaning* responders, subjects with a ratio of at least $33.\overline{3}$ %, but at most $66.\overline{6}$ % as *mixed-preference* responders, and all remaining subjects as *pragmatic-leaning* responders. Grouped like this, our post-exclusion sample of 100 participants can be said to consist of 37 literal-leaning, 43 mixed-preference, and 20 pragmatic-leaning responders.

Now, recall that one of the theoretical arguments made back in Section 2.2.4 on the basis of previous work had spelled out a possible prediction about expectable differences *within* a group of literal-leaning responders: Namely, the idea was that either print exposure or fluid intelligence (or both) would divide literal-leaning responders into two groups, with the group of higher-performing individuals generally providing slower literal responses on critical trials. This rather specific prediction was actually not addressed by any of our *planned* exploratory analyses reported above in Section 5.4. Therefore, we catch up on filling this gap here, in form of an unplanned exploratory test.

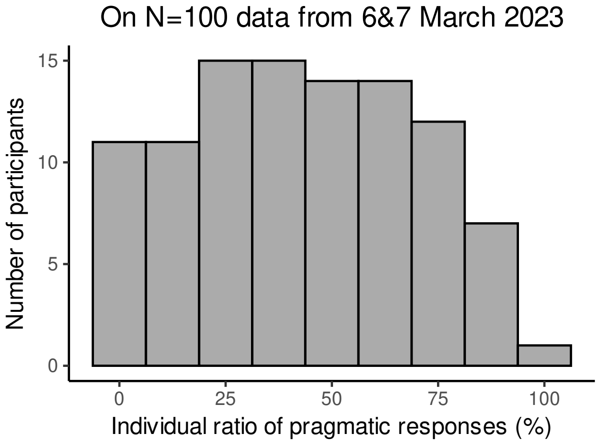


Figure 5.12: A histogram displaying participant-specific ratios of critical-trial responses that are pragmatic (given as percentages) in the data set of SPV responses collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion).

In order to test said prediction, we should again fit a linear mixed-effects model with residual log RT as its dependent measure, but this time just on a particular subset of observed responses to critical trials: The subset features only literal-leaning subjects and, also crucially, only the literal (but not any of the few pragmatic) responses that were given by such literal-leaning subjects. In the data we collected, subsetting based on this criterion results in a data set of 252 matching critical-trial observations. As predictors in our present model—which we may label Bonus Model—main effects of fluid intelligence, given as RPM scores, and print exposure, given as ART scores, are included as fixed effects, alongside by-subject and by-item random intercepts. Of course, the predictor variables fluid intelligence (a.k.a. `ID_GF`) and print exposure (a.k.a. `ID_PE`) need to be re-standardised now (mean to 0; SD to 1) with respect to the subset of data in question.

Bonus Model, only on critical trials literally verified by a literal-leaning responder:

- Gaussian distribution with identity link function
- `res_log_RT` $\sim$ `ID_GF + ID_PE + (1 | subject) + (1 | item)`

Results

Once fitting the Bonus Model to the relevant subset of data collected in March 2023 (reproduce by running script **osf.io/65chg** on data **osf.io/yd8mc** until line 1,034), the following fixed-effect estimates are obtained: While fluid intelligence does, indeed, appear to be a relevant predictor of literal response times among literal-leaning responders ($\hat{\beta}_{\texttt{ID\_GF}} = 0.08$, $SE = 0.03$, $t = 2.61$, $p = .013$) and, crucially, even in the hypothesised direction (higher intelligence $\sim$ greater slowdown), the analogous main-effect term of print exposure is found to be utterly irrelevant ($\hat{\beta}_{\texttt{ID\_PE}} = 0.02$, $SE = 0.03$, $t = 0.80$, $p = .429$). An overview of these fixed-effect estimates, including the one for the intercept of residual log RT ($\hat{\alpha}$), is provided in Table 5.4.

Now, given what we already know from the results in Section 5.4.3, i.e., of our *planned* exploratory model selections, the current pattern of results should not come as too much of a surprise: The fact that, again, a significant main effect of fluid intelligence, but no

Table 5.4: Fixed-effect estimates from the linear mixed-effects Bonus Model, fitted to residual log RT, based on SPV response data collected on 6 and 7 March 2023 from 37 literal-leaning Prolific participants (and considering only their literal responses). The included main-effect terms of fluid intelligence and print exposure are operationalised by performance scores in Raven's Progressive Matrices (RPM) and in an author recognition test (ART), respectively.

	$\hat{\alpha}$	SE	t	p
Intercept	0.33	0.08	4.41	.010
	$\hat{\beta}$	SE	t	p
Fluid intelligence (`ID_GF`)	0.08	0.03	2.61	.013
Print exposure (`ID_PE`)	0.02	0.03	0.80	.429

such main effect of print exposure is found when considering a subset of critical-trial observations (rather than the whole set) is quite expectable. So, we do not really obtain any ground-breaking news, here, regarding the properties of literal-leaning responders (and their literal responses) in particular since we already saw that the observed effect pattern is not unique to that group, but rather generally holds among all responder types (and both kinds of responses). Nonetheless, in an exploratory sense, it may be interesting to note anyway that the effect of fluid intelligence is *still* present in that subgroup of elevated interest, and actually even with a slightly higher estimate of magnitude, i.e., 0.08 log ms rather than just 0.06 log ms.

As for our theoretical considerations that were motivated by Tavano and Kaiser (2010) and were summarised back in Section 2.2.4, regarding how literacy ($\sim$ print exposure $\sim$ ART score) in particular should modulate a (literal-leaning) subject's verification response behaviour, there might be several ways to accommodate them with our present null finding concerning the main-effect term of ART score. Either they are simply wrong, or the assumed modulation is so subtle that we have too low statistical power to detect it, or one of the two linking hypotheses on the way from literacy to print exposure and then to ART score is misguided, just to name a few possibilities. In any case, on the present, shaky terrain of (unplanned) exploratory analysis endeavours, firm conclusions are better avoided.

5.5.2 Re-Analysis à la Van Tiel and Pankratz (2021)

As a further unplanned analysis, we choose to re-analyse our collected SPV response data using a method that closely mimics the approach undertaken by van Tiel and Pankratz (2021) in their paper. The goal is, again, to evaluate the predictions made by the polarity hypothesis regarding response-time latencies. That is, even though *both* our planned approach of fitting Model (i) in Section 6.3 *and* van Tiel and Pankratz' approach basically target that very same question, there are subtle differences. And in order for these differences not to get in the way of being able to interpret the conclusions of our hypothesis-driven analysis employing Model (i) in direct comparison to the previous findings by van Tiel and Pankratz, it would be good to show that they are not dependent on particularities of the chosen method of analysis.

Perhaps most strikingly, on the side of experimental design, it should be noted that van Tiel and Pankratz only ever show participants sentences employing the weaker term of an underlying Horn scale. For example, they would only ever show the sentence 'Red flowers are scarce.' to represent the scale ⟨scarce, absent⟩, in combination with any of three different picture variants displaying either few (unambiguous true), many (unambiguous false) or no red flowers (pragmatically ambiguous). That is, in their experiment, for each type of (ambiguous) critical trial there are two corresponding types of (unambiguous) control trials. By contrast, with our experimental design, we employ, e.g., both of the sentences 'She sometimes hit the bullseye today.' and 'She always hit the bullseye today.' in order to represent the scale ⟨sometimes $_{VBD}$, always $_{VBD}$⟩, each in combination with one of three picture variants. In consequence, for each type of critical trial we end up with five corresponding types of control trials (incl. three that are based on the stronger scalemate 'always $_{VBD}$'). Potentially, this inclusion of sentences with the stronger scalemate and thus the introduction of further and different kinds of control trials may have some unknown impact on the captured difference between the critical and the control condition overall. Therefore, for the analysis we describe here, we remove all data points stemming from trials where a sentence containing a stronger scalemate had to be responded to (i.e., with 'all', 'is definitely', 'always $_{VBD}$', 'everywhere $_{PP\text{-}LOC}$ it is', 'none', 'is definitely not', 'never $_{VBD}$', or 'nowhere $_{PP\text{-}LOC}$ is it') from our data set.

On that reduced data set, we now apply a linear mixed-effects model as specified right below using *lme4* (Bates, Mächler, Bolker and Walker, 2015), where `COND` represents condition (critical/control), `BIAS` the grand ratio of literal responses recorded for the given scale, `ORD` the index of a trial in the (individually randomised) order of appearance, and `scale` the given scale (here, not item!) among our overall eight Horn scales. The binary variables `COND`, `VR`, and `POL` are sum-coded, with positive signs for critical condition, 'True' response, and positive polarity. The continuous variables `BIAS` and `ORD` are centred at 0 and rescaled to display a standard deviation of 1. Let us refer to the model that is described here as VT&P'21-Like Omnibus Model. Unlike our previously reported models, this model is fitted relying on the restricted maximum likelihood (REML) criterion at convergence.

VT&P'21-Like Omnibus Model, on all trials featuring a weaker scalemate:

- Gaussian distribution with identity link function
- `log_RT ~ COND * VR * POL + BIAS + ORD + (1 | subject) + (1 | scale)`
- Custom optimisation settings:
  - `optimizer = 'optimx', calc.derivs = FALSE, method = 'nlminb', starttests = FALSE, kkt = FALSE`

This is precisely the model (incl. custom optimisation settings) that has originally been used by van Tiel and Pankratz (2021). Further, note that we consciously choose to include by-scale rather than by-item random effects here. Although this makes less sense in the context of *our* experimental design where scales are not entirely independent of each other, but rather are associated in a pair-wise manner (e.g., ⟨some, all⟩ to ⟨not all, none⟩), it is still useful to probe here in order to increase comparability to van Tiel and Pankratz and maybe even account for some scale-specific properties that are not captured by a by-item (i.e., by-pair-of-scales) random-effects structure. Since condition is now included directly as another predictor, instead of being used merely to residualise out some variation from the dependent measure, the fixed-effect term of interest is effectively the three-way interaction of `COND:VR:POL`, here. We may draw conclusions from this exploratory analysis, again as in Section 5.3, by computing t-values based on the Satterthwaite (1946) procedure for approximating degrees of freedom.

Second, van Tiel and Pankratz also conduct another analysis where they fit one model separately for each examined Horn scale. Then, they draw conclusions from the significance (and direction) of the `COND:VR` interaction in each of these scale-specific models (see 'Table 4' in their paper). Accordingly, we want to re-run an equivalent analysis on our data as well, not necessarily out of perceived necessity, but because it makes our results even easier to discuss in parallel to those reported by van Tiel and Pankratz:

VT&P'21-Like Scale-Specific Model (×8), for each scale, on all scale-specific trials featuring a weaker scalemate:

- Gaussian distribution with identity link function
- `log_RT ~ COND * VR + BIAS + ORD + (COND + VR || subject)`
- Custom optimisation settings:
  - `optimizer = 'optimx', calc.derivs = FALSE, method = 'nlminb', starttests = FALSE, kkt = FALSE`

Our implementation of this model type differs from van Tiel and Pankratz (2021) only in that we constrain random-effect intercorrelations to zero, here, which is necessary to allow the model to be supported by our given data. Note that the included predictor variables need to be re-coded for every particular scale-specific model so as to ensure that the resulting intercept always represents the estimated mean residual log RT in the particular scale-specific subset of data a model is fitted on. As stated, under this set-up, our fixed-effect term of interest has now become the `COND:VR` interaction. Once more, exploratory evaluations are conducted using the Satterthwaite (1946) procedure. For model fitting, the REML criterion at convergence is employed again, here.

Results

Based on our data collected in March 2023, we observe the following results (reproducible with script **osf.io/65chg** being run on data set **osf.io/yd8mc** until the very end, i.e., code line 1,134):

In the fitted VT&P'21-Like Omnibus Model, the three-way interaction of interest between condition, verification response, and polarity—which should come out significant according to the polarity hypothesis—is not significant ($\hat{\beta}_{\texttt{COND:VR:POL}} = -0.11$, $SE = 0.07$, $t = -1.64$, $p = .101$). This aligns well with our hypothesis-driven main result obtained from the same data, but using a slightly different statistical method (back in Section 5.3.6). Moreover, it is interesting to recognise the overall main effect of verification response ('True'/'False') here, which is quite marked and significant ($\hat{\beta}_{\texttt{VR}} = -0.06$, $SE = 0.02$, $t = -3.73$, $p < .001$) in the sense that people take considerably longer when providing a 'False' rather than a 'True' response (0.06 log ms, i.e., here, about 120 ms on average). It does seem quite plausible that participants would be faster in providing an affirmative ('True') rather than a rejecting ('False') response in general, due to the relative cognitive ease of affirming information rather than disaffirming it (for a discussion of this phenomenon in the context of SPV tasks, see Wang et al., 2021). There is also a very salient effect of order of trial presentation ($\hat{\beta}_{\texttt{ORD}} = -0.12$, $SE = 0.01$, $t = -15.77$, $p < .001$), indicating that, somewhat expectably, participants would steadily become faster in responding as they progressed through the individually randomised presentation of the overall 48 SPV trials, by about 0.12 log ms (i.e., here, roughly 250 ms) after every 14 trials (= the standard deviation of presentation order). Beyond that, we also see a significant main effect of condition (critical/control) where control trials seem to be processed and verified faster than critical trials ($\hat{\beta}_{\texttt{COND}} = 0.33$, $SE = 0.02$, $t = 20.95$, $p < .001$). This is something we have already established, in previous analyses (see Section 5.3.6), by interpreting the mean residual log RT as the additional delay associated with critical-trial responses. So, again, this just re-affirms observations already made previously relying on our own, residualisation-based analysis approach. Table 5.5 summarises all fixed-effect estimates obtained from the VT&P'21-Like Omnibus Model.

As for our eight VT&P'21-Like Scale-Specific Models, fitting them to the data leads us to observe the following: Only in two of them, namely those for the scale ⟨sometimes $_{VBD}$, always $_{VBD}$⟩ (+time) and the scale ⟨not everywhere $_{PP\text{-}LOC}$ is it, nowhere $_{PP\text{-}LOC}$ is it⟩ (−space), a significant interaction of interest between condition and verification response is found ($p < .05$). Both in the case of the +time scale ($\hat{\beta}_{\texttt{COND:VR}} = 0.24$, $SE = 0.09$, $t = 2.57$, $p = .011$) and in the case of the −space scale ($\hat{\beta}_{\texttt{COND:VR}} = 0.37$, $SE = 0.10$, $t = 3.64$, $p < .001$), this interaction has a positive sign. Given the way we have coded our variables, this translates to finding a *reversed* B&N effect for each of these two particular scales. By contrast, neither a (normal) B&N effect nor a reversed B&N effect is found for any of the six remaining scales; Table 5.6 illustrates that. Compare these findings to the polarity hypothesis' prediction that positively polar scales should consistently yield a B&N effect, while (explicitly) negatively polar ones should yield a reversed B&N effect. Evidently, there is quite a mismatch between that prediction and the given data, with the exception of the single scale −space that does, indeed, show a significant effect which also has the predicted sign. However, one must again be especially cautious not to interpret insignificant results (as obtained for six of the examined scales, here) as somehow demonstrating that there does not exist any real effect of either directionality: In such cases, it might very well be the case that there is, in fact, some real effect, but that it is too subtle in magnitude to be detectable with our current statistical power. A closer comparison to the study by van Tiel and Pankratz (2021) is quite enlightening in those regards: Their study, which at least found B&N effects among positive scales consistently, had collected 48 critical-trial observations from each of their (post-exclusion) 47 participants, i.e., 2,256 in total. Here, however, we have only collected 8 critical-trial observations from each of our (post-exclusion) 100 participants, i.e., 800 in total. Thus, our observed tendency to produce null findings with the present, scale-specific analyses is very possibly related to insufficient power. For comparability, an identical kind of analysis (see Section 6.4.2) is therefore repeated on a newly collected, four times larger response data set, gathered during our follow-up study that is now going to be presented.

Table 5.5: Fixed-effect estimates from VT&P'21-Like Omnibus Model, fitted to log RT, based on SPV response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). This table has a twin further below in Section 6.4.2 where analogous results from a follow-up study with larger sample size are discussed: Table 6.5.

	$\hat{\alpha}$	*SE*	*t*	*p*
Intercept	7.63	0.05	155.48	< .001
	$\hat{\beta}$	*SE*	*t*	*p*
Condition (`COND`)	0.33	0.02	20.95	< .001
Verification response (`VR`)	−0.06	0.02	−3.73	< .001
Polarity (`POL`)	−0.18	0.09	−2.10	.090
Response bias (`BIAS`)	0.05	0.04	1.23	.274
Order of presentation (`ORD`)	−0.12	0.01	−15.77	< .001
Interaction of `COND:VR`	0.04	0.03	1.07	.286
Interaction of `COND:POL`	−0.00	0.03	−0.05	.962
Interaction of `VR:POL`	−0.03	0.03	−0.87	.387
Interaction of `COND:VR:POL`	−0.11	0.07	−1.64	.101

Table 5.6: Estimates for the fixed effect of the interaction between condition and verification response, from each of the eight VT&P'21-Like Scale-Specific Models, fitted to log RT, based on SPV response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). The rightmost column summarises if the effect of interest is either insignificant ('—') at $p < .05$, resembling a B&N effect ('present'), or resembling a *reversed* B&N effect ('reversed'). This table has a twin further below in Section 6.4.2 where analogous results from a follow-up study with larger sample size are discussed: Table 6.6.

Scale	Interaction of `COND:VR`				
	$\hat{\beta}$	*SE*	*t*	*p*	B&N
+quantity	0.01	0.10	0.09	.929	—
+possibility	−0.25	0.13	−1.96	.051	—
+time	0.24	0.09	2.57	.011	reversed
+space	0.20	0.11	1.84	.068	—
−quantity	0.15	0.10	1.40	.163	—
−possibility	−0.05	0.09	−0.58	.565	—
−time	0.02	0.11	0.17	.863	—
−space	0.37	0.10	3.64	< .001	reversed

Chapter 6
Study Two

In this chapter, we describe a second experimental study that was conducted based on the experimental materials and design presented in Chapter 4. Crucially, as opposed to Study One (Chapter 5), this present follow-up study—let us call it *Study Two*—administered only one single task to participants, the SPV task. The experiment for Study Two was carried out via the crowd-sourcing platform Prolific on 6 April 2023. We present the underlying motivation for this study in Section 6.1 and then move on to describing two preparatory steps undertaken prior to conducting the study in Section 6.2, i.e., a Bayesian meta-analysis of various previous studies as well as a simulation-based power analysis used to determine the prespecified sample size. In Section 6.3, the method behind our main, hypothesis-driven analysis employed here is delineated (as preregistered at **osf.io/dhpzq**), alongside the results eventually obtained by applying it to the gathered experimental data. Afterwards, Section 6.4 presents two planned, but exploratory re-analyses of the same data using two different approaches (both of them preregistered at **osf.io/dhpzq** as well), i.e., a Bayes factor analysis and, again for comparability purposes, the same statistical method as van Tiel and Pankratz (2021). At last, another, more comprehensive way of statistically modelling our collected data is reported in Section 6.4 in form of an unplanned exploratory analysis.

6.1 Motivation

As has been shown (Section 5.3.6), the two hypothesis-driven analyses in Study One both returned null results. One of these two analyses, employing the statistical model we labelled Model (i), was designed to address the polarity hypothesis (Section 2.1.3) about polarity-sensitive differences in response-time latencies during SI processing, while the other one, employing Model (ii), addressed a secondary, polarity-based hypothesis (Section 2.1.4) whose dependent measure of interest were the verification response choices themselves that participants made.

Our exclusive concern, now, is to test the *polarity hypothesis* again (however not our secondary hypothesis), through conducting a properly powered follow-up study. This is necessary for us to be able to more reliably accept or dismiss the polarity hypothesis on the basis of our eventual findings, especially considering that van Tiel and Pankratz (2021), who defend the polarity hypothesis, have collected a data set consisting of a much larger amount critical-trial observations than we have done in Study One, as was discussed towards the end of Section 5.5.2.

6.2 Preparation

During the process of designing Study Two, we conducted a simulation-based power analysis (Section 6.2.1) and a Bayesian meta-analysis of previous studies (Section 6.2.2) which we both report here. Note that by the end of Section 6.2.1 one can also find a brief description of the pre-exclusion sample of 432 recruited Prolific participants, alongside a summary of demographic properties of the post-exclusion subsample of $N = 400$ such participants who were eventually considered in the statistical analyses of Study Two.

6.2.1 Power Simulations to Determine Sample Size

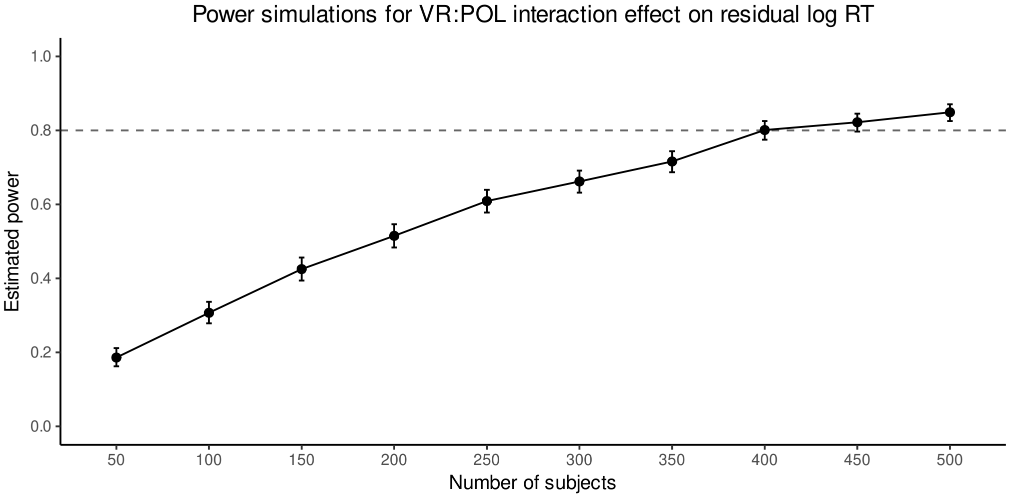


Figure 6.1: Estimated power for detecting `VR:POL` interaction effect (i.e., between verification response and polarity) on residual log RT, computed in order to determine a planned sample size for Study Two, probing different possible sample sizes from 50 to 500. The underlying, assumed effect-size estimate is drawn from response data from overall 135 subjects, collected between 16 February and 7 March 2023. 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.

As just stated, for the present follow-up study, the planned post-exclusion number of participants is $N = 400$. This is motivated by the results of another prospective, simulation-based power analysis that we conducted (reproduce with script **osf.io/zypeq** on data **osf.io/bu9wa**), again implemented using the *simr* package (Green and MacLeod, 2016) for the R programming language (R Core Team, 2023). That power analysis relied on effect-size estimates drawn from the entire body of response data from (the final version of) our SPV task that we had collected so far. That is, they were computed based on a combined data set of the 16-February pilot study ($N = 20$), the 6-and-7-March actual study, i.e., Study One ($N = 100$), and what we may sloppily call ‘leftover’ data which was also collected on 6 and 7 March 2023 ($N = 15$).[21] Hence, this is a power analysis that is more to be trusted in its accuracy than the earlier one we reported in Section 5.3.5, considering that now we can rely on a much larger corpus of data to draw a realistic effect-size estimate from. In fact, the VR:POL effect size of interest as calculated on the given overall sample of 135 people ($\hat{\beta}_{\texttt{VR:POL}} = -0.09$, on the log-milliseconds scale) is slightly closer to zero than in the subsample of 100 people analysed for the main purpose of our previous study (cf. back in Section 5.3.6). As it turns out, a new sample size of at least 400 is needed to reach statistical power of 0.8 (our desired threshold) for finding a real such interaction effect of the assumed magnitude, given that 801 out of 1,000 simulations were successful for this particular sample size. The full results of the present power analysis are illustrated by the plot in Figure 6.1. Consequently, we opt for a post-exclusion sample size of $N = 400$ for the experiment of Study Two.

Relying on the stopping rule derived from this planned post-exclusion sample size, we ended up recruiting 432 Prolific workers in total (pre-exclusion). Actually, we could (and should) have already stopped after recruiting 429 Prolific workers in order to reach post-exclusion $N = 400$, but we accidentally went even slightly beyond that (by three subjects) due to an interim miscalculation of how many subjects had already been excluded based on their accuracy on SPV control trials, caused by a floating-point error. Hence, we did not consider the data of the three chronologically very last participants who joined the study.[22] Among the remaining 429 subjects, there was one who revoked their consent to participation after completing the experiment; their data is, thus, not analysed here. Two subjects experienced a rare technical glitch that seemingly had them skip through the entire SPV task without actually getting to respond to any trials. Then, 26 further subjects were regularly excluded based on the prespecified criterion of not meeting the 80 % accuracy threshold in the control condition of the SPV task (see Section 4.2). This is how we eventually end up with our post-exclusion sample that consists of $N = 400$ participants.

Note that these 400 participants display the following summary demographics: Their mean age at the time they took the study was 35 years, with a sample SD of 3 years. The minimum age was at 30 and the maximum age at 40 because sampling was constrained as described in Section 4.2. There were 241 men (60.25 %), 150 women (37.5 %), and 9 people of non-binary gender (2.25 %) among the subjects (cf. biological sex: 245 males [61.25 %], 154 females [38.5 %], 1 undisclosed [0.25 %]). While all subjects were native speakers of English, 51 (12.75 %) of them had self-reportedly grown up as bilinguals, i.e., with more than one language.

6.2.2 Meta-Analysis of Previous Studies

While preparing Study Two, we also carried out a Bayesian meta-analysis of the results of seven experiments reported in previous literature regarding the magnitude of the VR:POL interaction effect on residual log RT, conjointly with estimates drawn from our own experimental data that we had collected so far. That is, apart from our own data, seven experimental results reported in (or indirectly retrievable from) the following five publications were considered: Cremers and Chemla (2014), Bill et al. (2018), van Tiel et al. (2019), Marty et al. (2020), and van Tiel and Pankratz (2021). What all these

[21] This encompasses responses from participants who did meet our SPV inclusion criterion, but failed to meet some other inclusion criterion based on one of the individual-differences tasks or aborted the experiment at some point, be it voluntarily or due to some technical issue.

[22] For transparency purposes, the response data from these three disconsidered 'spillover' participants is stored separately here: **osf.io/ujdr2**.

Table 6.1: Studies or data sets included in the meta-analysis. Their respective post-exclusion participant sample sizes are listed in the column labelled N. Estimates for the mean residual log RT (i.e., intercept estimates $\hat{\alpha}$) and the magnitude of the interaction of interest (i.e., slope estimates $\hat{\beta}_{\texttt{VR:POL}}$) are listed in the next two columns. In brackets next to each such estimate, a standard error is reported. Information on whether these estimates were *computed* directly by (re-)analysing data using the method described in Section 5.3 or if they were rather *approximated* from summary statistics and (for the standard errors) by comparative imputation, all due to a lack of publicly accessible data, is provided in the rightmost column.

Study / Data	N	$\hat{\alpha}$	$\hat{\beta}_{\texttt{VR:POL}}$	Acquisition
Cremers and Chemla (2014), Exp. 1	36	0.26 (0.05)	−0.49 (0.10)	Approximated
Cremers and Chemla (2014), Exp. 2	60	0.13 (0.04)	0.00 (0.08)	Approximated
Bill et al. (2018)	35	0.39 (0.05)	−0.00 (0.10)	Approximated
Van Tiel et al. (2019), Exp. 1	49	0.16 (0.03)	−0.47 (0.07)	Computed
Van Tiel et al. (2019), Exp. 3	371	−0.04 (0.03)	−0.18 (0.04)	Computed
Marty et al. (2020)	147	0.16 (0.03)	−0.69 (0.05)	Approximated
Van Tiel and Pankratz (2021)	47	0.18 (0.02)	−0.29 (0.07)	Computed
Our 16-February Pilot Study	20	0.28 (0.06)	−0.37 (0.14)	Computed
Our 6&7-March Actual Study	100	0.37 (0.05)	−0.10 (0.07)	Computed
Our 6&7-March Leftover Data	15	0.26 (0.07)	0.15 (0.16)	Computed

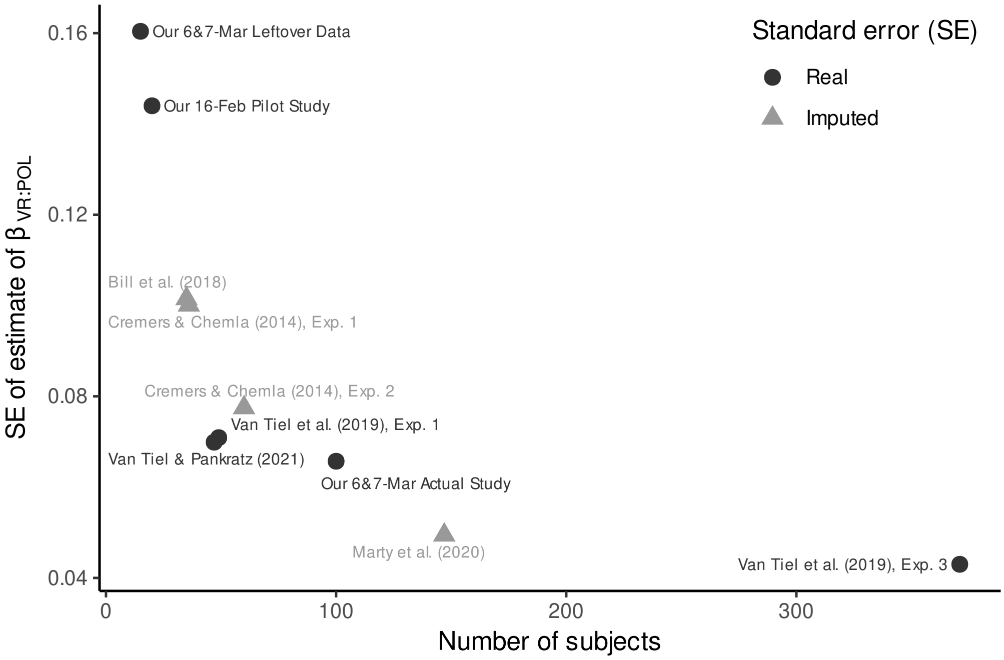


Figure 6.2: Standard errors (SEs) computed for estimates of the interaction of interest for the data sets resulting from studies by van Tiel et al. (2019), by van Tiel and Pankratz (2021), and by ourselves are plotted as black dots. The horizontal axis represents the (post-exclusion) number of subjects N of any particular study or data set. Additionally, exploiting the fact that, for these computed (i.e., real) SE values, the relationship between SE and $1/\sqrt{N}$ is roughly linear, imputed standard errors for the remaining studies are plotted as grey triangles, as predicted by a simple linear model fitted to capture that relationship.

experiments had in common was that they required participants to verify sentences that displayed pragmatic ambiguity triggered both by positively polar scales and (in other trials/conditions) by negatively polar scales. That being said, of course there were also marked differences across the designs of these experiments, as already discussed to some extent back in Section 2.1.5: Cremers and Chemla asked participants to verify sentences against their world knowledge (e.g., 'Some elephants are mammals.'), whereas van Tiel et al., van Tiel and Pankratz, and we ourselves asked them to verify sentences against pictures. Bill et al.'s experiment also employed pictures, but within a Covered Box (Huang et al., 2013) paradigm rather than a classical SPV (Clark and Chase, 1972) paradigm. Furthermore, Cremers and Chemla, Bill et al., Marty et al., and we ourselves compared polarity within symmetrical pairs of scales, whereas van Tiel et al. and van Tiel and Pankratz compared some positive scales against some negative scales without their being a symmetrical, pair-wise relation between the scale sets of either polarity value. Lastly, while the majority of experiments relied on participants' intuitive verification judgements, Cremers and Chemla (2014), Exp. 2 and van Tiel et al. (2019), Exp. 3 are exceptions to this as they trained half of their participants beforehand to respond literally and another half to respond pragmatically. In spite of these differences across studies, however, we can still compare their respective effect-size estimates for the interaction of interest, under the assumption that there exists a latent effect of the kind proposed by the polarity hypothesis 'out there in the world' which should be robust against the moderate variation caused by such superficial discrepancies.

To the best of our knowledge, there are no other published studies that specifically compare response times in SI processing between positively and negatively polar scales beyond those reported in the five publications we have mentioned.[23] Where data was openly accessible (van Tiel et al., 2019; van Tiel and Pankratz, 2021; and obviously our own data) we computed estimates for mean residual log RT (i.e., an intercept estimate $\hat{\alpha}$) and for the interaction of interest (i.e., a slope estimate $\hat{\beta}_{\texttt{VR:POL}}$) using our frequentist modelling method described back in Section 5.3. This way, we also calculated the standard errors (SEs) for both of these two estimates. Where data was not accessible, we approximated such estimates from close and careful inspection of all summary statistics, including plots, in the papers themselves. Regarding the approximation of corresponding SEs, we resorted to imputing them from the roughly proportional relationship between the inverse of the square root of the number of subjects and the standard error (for either $\hat{\alpha}$ or $\hat{\beta}_{\texttt{VR:POL}}$), as it was observed across the few data sets that were, in fact, available to us. That is, separately for intercept SEs and interaction-slope SEs, we fitted a simple linear model predicting SE from the inverse of the square root of the number of (post-exclusion) participants. Accordingly, we then computed predicted SE values for each of the experiments whose data we did not have based on its reported (post-exclusion) participant sample size. Visually, and just for the interaction-slope SEs, this imputation process can be acknowledged in Figure 6.2. All this gives rise to the summary of intercept and interaction-slope estimates, alongside standard errors, reported in Table 6.1. Next, we implemented and fitted the following two Bayesian hierarchical linear models in *brms* (Bürkner, 2017) in order to generate a meta-analysis posterior distribution, respectively, for $\hat{\alpha}$*'s* as well as for $\hat{\beta}_{\texttt{VR:POL}}$'s, using the prior $N(0,\,1)$ for all model parameters:[24]

[23] One may ask why Romoli and Schwarz (2015) is not also considered, recalling that it is one of the studies that van Tiel and Pankratz (2021) themselves cite as providing evidence that is consistent with (and therefore relevant to) the polarity hypothesis. That is because the experiment conducted by Romoli and Schwarz actually did not feature any positively polar Horn scales at all (only a negatively polar one). Hence, there is no `VR:POL` interaction magnitude to be measured based on their results.

[24] $N(0,\,1)$ refers to a normal distribution with mean 0 and standard deviation 1.

Meta-Analysis Model of $\hat{\alpha}$'s, on summary data from Table 6.1:

- Gaussian distribution with identity link function
- `alpha_hat | resp_se(SE_alpha_hat, sigma = FALSE)` $\sim$ `1 + (1 | study)`
- Custom optimisation settings:
  - `adapt_delta = 0.99, max_treedepth = 10`

Meta-Analysis Model of $\hat{\beta}_{\texttt{VR:POL}}$'s, on summary data from Table 6.1:

- Gaussian distribution with identity link function
- `beta_hat | resp_se(SE_beta_hat, sigma = FALSE)` $\sim$ `1 + (1 | study)`
- Custom optimisation settings:
  - `adapt_delta = 0.99, max_treedepth = 10`

As is reproducible by running the R script **osf.io/85ht2**, the resulting meta-analysis posterior for the intercept estimates ($\hat{\alpha}$'s) resembles a $N(0.212, 0.050)$ distribution, while the posterior for the interaction-slope estimates ($\hat{\beta}_{\texttt{VR:POL}}$'s) resembles a $N(-0.232, 0.107)$ distribution. This is useful information as it now can tell us what effect magnitude for the interaction of interest we should believe to expect in future data, given all prior evidence, i.e., a magnitude of -0.232 log ms (with a 95 % credible interval of $[-0.422, -0.022]$), and also what mean additional delay associated with critical trials we are likely to witness, i.e., a mean residual log RT of 0.212 log ms (with a 95 % credible interval of $[0.114, 0.310]$). Moreover, we are going to employ the posteriors resulting from this meta-analysis as priors in a Bayes factor (re-)analysis of the eventually collected $N = 400$ SPV response data from the present Study Two, as reported later on in Section 6.4.1.

6.3 Single Planned Hypothesis-Driven Analysis

Here, in Study Two, there is only a single planned hypothesis-driven analysis to be conducted on the newly collected response data from the SPV task. We summarise its underlying method very concisely right below in Section 6.3.1 as it is actually identical to an earlier analysis carried out previously in Study One. Then, we report the obtained results in Section 6.3.2.

6.3.1 Summary of Method

We are working with the predictor variables verification response (`VR`) and polarity (`POL`) again. Each of them is sum-coded with regard to the data set of critical trials. In doing so, literal ('True') responses and positive polarity are respectively coded as values with a positive sign. For more details, please check the analogous section from Study One on predictor variables (Section 5.3.2).

Our dependent measure of interest is residual log RT (`res_log_RT`), encoding the variation in log RT that is unique to critical trials. For details on why this particular measure is constructed, please revisit Sections 5.3.2 and 5.3.3 from Study One.

The analysis steps about to be described are carried out using the *lme4* package (Bates, Mächler, Bolker and Walker, 2015) for R (R Core Team, 2023). First, using the *bobyqa* optimiser, we fit the following linear mixed-effects model—labelled New Control Model—to observed log response times (`log_RT`) during control trials:

New Control Model, on control trials:

- `log_RT ~ VR * POL + (VR * POL | subject) + (VR * POL | item)`
- Non-convergence (incl. singular fit): Remove by-item `VR:POL` slope, then by-subject `VR:POL` slope, then by-item `VR` slope, then by-item `POL` slope, then by-subject `VR` slope, then by-subject `POL` slope, keep intercepts.

As before, we then use the fitted Control Model to generate predicted values of log RT within the data set of critical trials. For each critical-trial observation, we compute its residual log RT by subtracting the control-predicted value from the observed log RT. Next, again using the *bobyqa* optimiser, we fit the following (also frequentist) linear mixed-effects model to residual log RTs during critical trials and call it New Model (i):

New Model (i), on critical trials:

- `res_log_RT ~ VR * POL + (VR * POL | subject) + (VR * POL | item)`
- Non-convergence (incl. singular fit): Remove by-item `VR:POL` slope, then by-subject `VR:POL` slope, then by-item `VR` slope, then by-item `POL` slope, then by-subject `VR` slope, then by-subject `POL` slope, keep intercepts.

Relying on the Satterthwaite (1946) procedure for approximating degrees of freedom, we decide to carry out one null-hypothesis significance test on the New Model (i), aimed at probing the relevance of the fixed-effect `VR:POL` interaction term. As our threshold for interpreting statistical significance, we set $p < .05$.

If the polarity hypothesis is true, a significant `VR:POL` interaction effect should be found, and notably with a negative-sign parameter estimate $\hat{\beta}_{\texttt{VR:POL}}$. Despite this directional expectation, we carry out a two-sided statistical test again, i.e., one that would be sensitive to significantly non-zero interaction-slope estimates $\hat{\beta}_{\texttt{VR:POL}}$ of either sign.

6.3.2 Results

In the following, we present the results obtained from the one hypothesis-driven analysis that was performed on Study Two's SPV response data from 6 April 2023. It can be reproduced by running the script stored in **osf.io/ve9ab** on the data in **osf.io/y4pkm** until line 415.

Table 6.2: Fixed-effect estimates from linear mixed-effects New Model (i), fitted to residual log RT, based on SPV response data collected on 6 April 2023 from 400 Prolific participants (post-exclusion). Corresponding random-effect estimates from the same model are summarised further below in Appendix C, in Table C.1. Moreover, this table has a twin further above in Section 5.3.6 where analogous results from our previous study with smaller sample size are discussed: Table 5.1.

	$\hat{\alpha}$	*SE*	*t*	*p*
Intercept	0.34	0.07	4.88	.008
	$\hat{\beta}$	*SE*	*t*	*p*
Verification response (`VR`)	−0.04	0.02	−2.40	.016
Polarity (`POL`)	−0.08	0.02	−4.91	< .001
Interaction of `VR:POL`	−0.08	0.03	−2.28	.023

The interaction of interest between verification response and polarity does, indeed, show a significant effect here in Study Two ($\hat{\beta}_{\texttt{VR:POL}} = -0.08$, $SE = 0.03$, $t = -2.28$, $p = .023 < .05$). Importantly, the estimated effect even carries the predicted sign ($\hat{\beta}_{\texttt{VR:POL}} < 0$). Its absolute magnitude is approximately 0.08 log ms, which translates to about 135 ms when comparing observations that are close to the average response time found in the data set. This result seems to provide support for the polarity hypothesis.

Table 6.2 summarises all fixed-effect estimates derived from this analysis, i.e., by fitting the New Model (i). Beside the fixed-effect parameter of interest which was just discussed, it is also interesting to observe, in an exploratory manner, that the main effects both of verification response ($\hat{\beta}_{\texttt{VR}} = -0.04$, $SE = 0.02$, $t = -2.40$, $p = .016$) and of polarity ($\hat{\beta}_{\texttt{POL}} = -0.08$, $SE = 0.02$, $t = -4.91$, $p < .001$) on residual log RT display significant nonzero slopes as well. In case of verification response, this is to say that pragmatic responses are generally associated with longer residual log RTs than literal responses. For polarity, the observed pattern is that negatively polar trials are generally associated with longer residual log RTs than positively polar trials. A visualisation of these general relationships is brought forth by the plot in Figure 6.3. Lastly, the intercept for residual log RT is estimated to be at 0.34 log ms ($\hat{\alpha} = 0.34$, $SE = 0.07$, $t = 4.88$, $p = .008$). Again, thanks to the employed variable coding scheme, we can interpret it as the average residual log RT and, hence, as the average additional delay associated with critical trials when compared to control trials.

Note that a complementary overview of random-effect parameter estimates from the fitted New Model (i) is provided in Appendix C, in Table C.1.

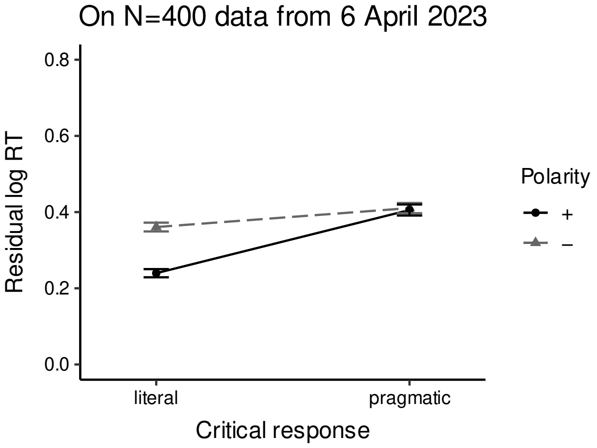


Figure 6.3: Mean residual log RTs grouped by (critical-trial) verification response and by polarity. The displayed error bars represent within-subject-and-item standard errors (computed following Morey et al., 2008). The underlying SPV response data was collected on 6 April 2023 from 400 Prolific participants (post-exclusion). One is invited to compare this plot to a sister plot in Appendix C (Figure C.1) where log RTs rather than residual log RTs are displayed based on the same data. Moreover, there is a twin plot further above in Section 5.3.6 where analogous results from our previous study with smaller sample size are presented: Figure 5.2.

Speaking of random effects, the random-effects structure of our convergingly fitted variant of New Model (i) features only by-subject and by-item random intercepts, but no random slopes whatsoever. By contrast, the selected variant of the New Control Model, initially fitted on control trials in order to allow for residualisation as discussed, features not only by-subject and by-item random intercepts, but also by-subject random slopes for verification response and polarity as well as by-item slopes for (just) polarity; it also includes all accordingly possible intercept–slope or slope–slope correlation terms as further random parameters. Recall that, for either model, the random-effects structure that was eventually opted for was chosen strictly following our (maybe suboptimal) preregistered, deterministic selection criterion described in Section 6.3.1.

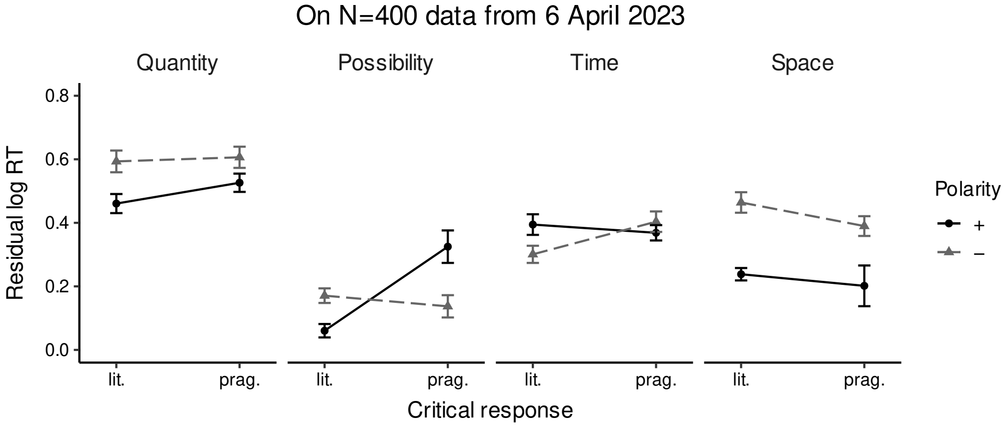


Figure 6.4: For each item separately: Mean residual log RTs grouped by (critical-trial) verification response and by polarity. The displayed error bars represent on within-subject standard errors (computed following Morey et al., 2008). The underlying SPV response data was collected on 6 April 2023 from 400 Prolific participants (post-exclusion). One is invited to compare this plot to a sister plot in Appendix C (Figure C.2) where log RTs rather than residual log RTs are displayed based on the same data. There is also a twin plot further above in Section 5.3.6 where analogous results from our previous study with smaller sample size are presented: Figure 5.3.

Now, it may be worth comparing the random-effects structures we ended up with, here, with those analogous ones from Study One so as to highlight some potential methodological problems: Back in Study One (see Section 5.3.6), it was the Control Model that featured the more simple structure `(1|subject)+(1|item)`, whereas Study One's Model (i) converged with the more complex structure `(VR+POL|subject)+(POL|item)`. So, as can be noticed now, the exact opposite is the case here in Study Two, with the New Control Model featuring the structure `(VR+POL|subject)+(POL|item)`, but where the New Model (i) converges with the structure `(1|subject)+(1|item)` instead. In consequence, we are actually dealing with two subtly different things when referring to 'residual log RT' in Study One versus referring to 'residual log RT' in Study Two: In case of Study One, we are effectively talking about a measure for critical trials derived by residualising out population-level variation patterns caused by effects of verification response and polarity that are also present in control trials as well as group-level differences caused by higher/lower average speeds of responding to control trials for particular subjects/items. But here, in case of Study Two, we are rather talking about a measure derived by residualising out *not only* the kind of variation just described, *but also* such group-level control variation that can be modelled by taking into account steeper/milder slopes of verification response or polarity for particular subjects/items. This intricate discrepancy could then also be the reason why Model (i) from Study One is supported by the data even with rather complex random effects, whereas New Model (i) from Study Two converges only with simpler random effects: Since the dependent measure of each is a slightly different kind of 'residual log RT', there is now one case—the case of Study One—in which only intercept-based random variation has already been removed from the measure of interest, hence leaving aside large amounts of *slope-based* group-level variation that can still be modelled by Model (i), but there is also another case—Study Two—where rather complex patterns of random variation (incl. slope-based ones) have 'already been dealt with' during residualisation and thus 'leave nothing much behind to be explained' by potentially added random slopes in New Model (i), hence 'forcing' New Model (i) into a simpler random-effects structure in order to reach convergence. Although this issue might seem rather minor at first glance, it does call for a closer inspection of it on more formal grounds (i.e., beyond the merely conceptual/intuitive argument just made) in order to get a better understanding of its consequences and if these are desirable, undesirable, or, at least, irrelevant to our main purpose of drawing a meaningful statistical inference about the fixed-effect parameter `VR:POL` in each of our two present experimental studies. On a final note, it may also be useful to think about this issue as an element within a vaster array of problems associated with various kinds of residualisation-based analysis approaches (for a discussion, see York, 2012).

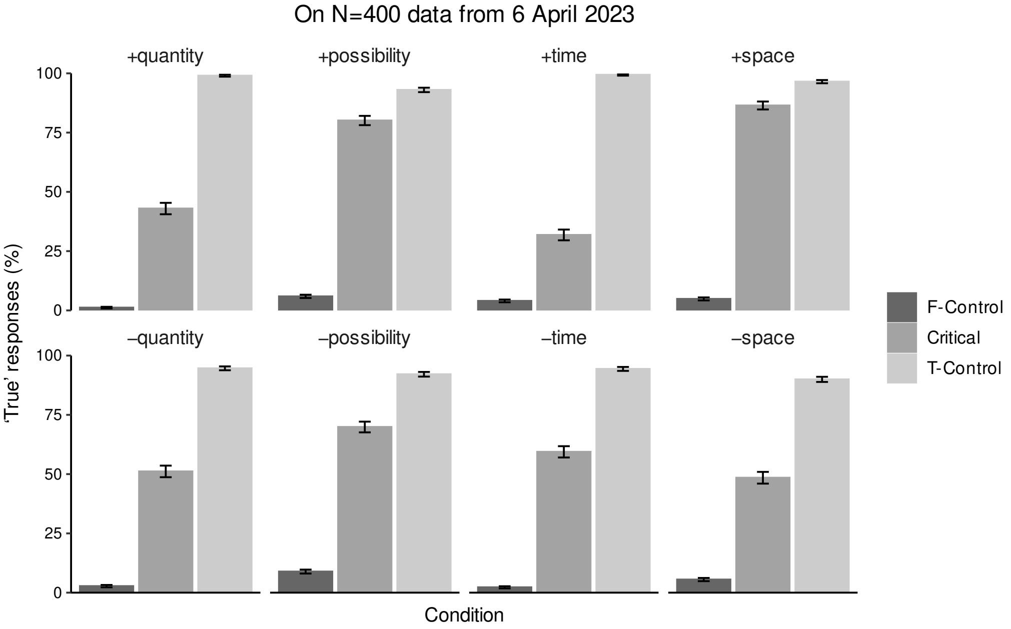


Figure 6.5: Bar plots showing ratios of 'True' responses in the SPV data collected on 6 April 2023 from 400 Prolific participants (post-exclusion), grouped by scale (+quantity, +possibility, ..., −space) and by three different condition set-ups: control trials where the expected response is 'False' (F-Control), critical trials (Critical), and control trials with a 'True' expected response (T-Control). The displayed error bars represent within-subject standard errors (computed following Morey et al., 2008). Note that there is a twin figure further above in Section 5.3.6 where analogous results from our previous study with smaller sample size are presented: Figure 5.6.

Closing this bracket, let us return to the plotted mean residual log RTs in Figure 6.3 and see how they compare to those in a further plot given in Figure 6.4 where the data has additionally been grouped by item. Quite surprisingly, we see that only the Possibility item displays a very obvious, 'X'-shaped effect pattern as would be 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). Thus, it seems that the overall significant effect of the VR:POL interaction that we find using our hypothesis-driven test is mainly 'carried' by the unusually salient effect pattern among the Possibility item's trials. Indeed, if one were to run the exact same analysis (exploratorily) on a subset of data without any trials of the Possibility item, the effect would actually vanish completely (without Possibility item: $\hat{\beta}_{\texttt{VR:POL}} = 0.01$, $SE = 0.04$, $t = 0.23$, $p = .820$). Even more problematically, we observe that the +possibility scale has one of the highest ratios of literal responses (and thus the lowest of pragmatic responses) associated with it, similarly as also found in the response data from Study One. In fact, only 19.9 % of observed responses to +possibility critical trials were pragmatic. At the same time, it is precisely this small set of 'unusual' responses to +possibility critical trials that comes with a markedly higher average residual log RT (i.e., one of 0.32 log ms) than any of the three further combinations of response type and polarity within the Possibility item (between 0.06 and 0.17 log ms). Hence, it is easy to see how this particular group of responses is the most crucial one in accounting for the overall VR:POL interaction effect detected by our hypothesis-driven test. But does this group of responses necessarily reflect genuine, on-line pragmatic processing that should be of interest here? It is possible, at least, to speculate if a rare and, thus, 'unusual' kind of response to a particular type of experimental trial might not rather reflect some sort of conscious response strategy adopted by a salient minority of participants, who have concluded that the modal-verb expression 'might be' is desired *by the experimenter* to be treated in an analogous way to, e.g., the quantifier pronoun 'some' which also appears in the very same experiment, but for which drawing pragmatic inferences is actually a much more common and, thus, 'natural' thing to do. A conscious response strategy of this kind, aimed at delivering what is perceived to be consistent response behaviour as desired by the experimenter, might then account for the longer response delay associated, on average, with this particular group of observations. Find a comprehensive overview of literal-response ratios by condition in Figure 6.5.

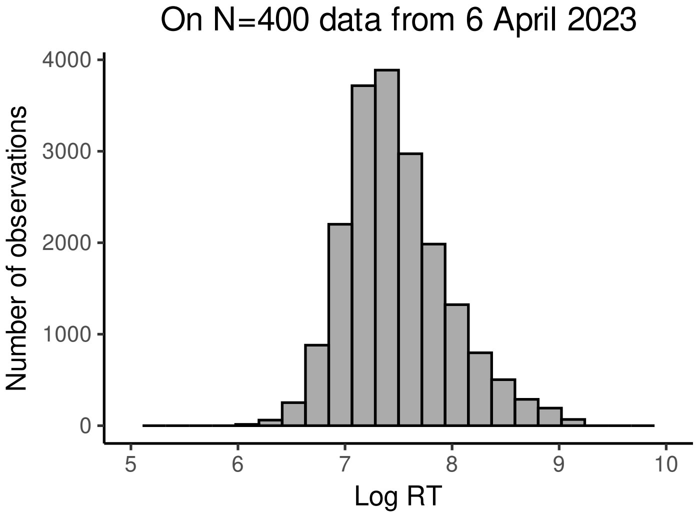


Figure 6.6: A histogram that shows the distribution of log RTs in the full (critical + control) data set of SPV response data collected on 6 April 2023 from 400 Prolific participants (post-exclusion). Here, log RT stands for the natural logarithm of a raw response time given in milliseconds. There is a twin plot further above in Section 5.3.6 where analogous results from our previous study with smaller sample size are presented: Figure 5.4.

Finally, Figures 6.6 and 6.7 display histograms of the distributions of log RTs and of residual log RTs, respectively, as found in our newly collected experimental data from Study Two. Again, they can serve as visual sanity checks that these measures are distributed in a way that is not too divergent from a normal, i.e., Gaussian distribution, thus congruing well with one of the relevant model assumptions behind each of the two linear mixed-effects models that have been fitted here.

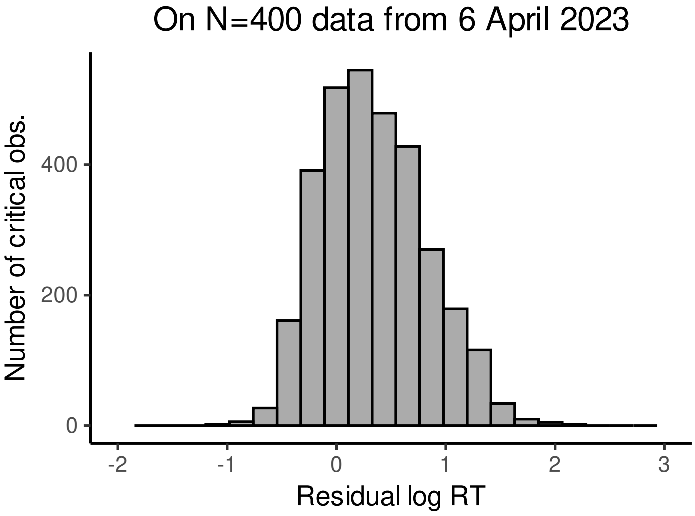


Figure 6.7: A histogram that shows the distribution of residual log RTs in the data set of critical-trial SPV response data collected on 6 April 2023 from 400 Prolific participants (post-exclusion). Residual log RT is calculated by subtracting a New-Control-Model-predicted value for log RT from the observed log RT of a critical trial (see Section 6.3.1). In this context, log RT stands for the natural logarithm of a raw response time given in milliseconds. There is a twin plot further above in Section 5.3.6 where analogous results from our previous study with smaller sample size are presented: Figure 5.5.

6.4 Planned Exploratory Analyses

The following two subsections cover two alternative ways of analysing the newly collected SPV response data, but fundamentally still with the same question in mind as before in our main analysis, i.e., the question whether the polarity hypothesis holds true. The first approach is a Bayes factor analysis, evaluated while comparing different prior specifications, and the second one is an alternative (frequentist) re-analysis of the data which proceeds exactly along the lines of van Tiel and Pankratz (2021) as we have already done previously for Study One's data. Here, both of these additional analyses had been committed to and registered prior to data collection.

6.4.1 Re-Analysis Using Bayes Factors

In this section, we describe how we re-analyse the SPV response data that was collected for Study Two using the Bayesian statistical framework. We do this by fitting a linear mixed-effects (i.e., hierarchical) model very similar to the frequentist one that our main analysis (Section 6.3) relied on. Crucially, however, we now probe how three differentially informative prior specifications each modulate the resulting posterior distribution. Using

Table 6.3: An overview of prior specifications that are probed in the present Bayes factor analysis. A probability-density plot of the three different priors for $\beta_{\texttt{VR:POL}}$ is provided in Figure 6.8.

Prior specification	α	$\beta_{\texttt{VR:POL}}$
Uninformative	$N(0,\, 0.5)$	$N(0,\, 0.5)$
Meta-analysis posterior	$N(0.212,\, 0.050)$	$N(-0.232,\, 0.107)$
Lik. est. from only our data	$N(0.349,\, 0.048)$	$N(-0.093,\, 0.056)$

a Bayes factor analysis, we then test how each of the three eventually fitted models competes against its corresponding null model which assumes the effect size of the `VR:POL` interaction of interest to be zero.

Table 6.3 lists three different prior specifications for the intercept (α) a.k.a. the mean residual log RT and the `VR:POL` slope ($\beta_{\texttt{VR:POL}}$) a.k.a. the effect size of the interaction of interest on the log-milliseconds scale. The first one is quite straightforward: It features a normal prior distribution with mean zero and standard deviation 0.5 for both α and $\beta_{\texttt{VR:POL}}$, which covers a broad range of plausible values on the log-milliseconds scale and is therefore relatively uninformative. The second one features the posteriors of the meta-analysis we conducted and reported in Section 6.2.2 as priors. Lastly, the third one features rather tight priors that represent likelihood estimates drawn from (only) our own combined previous data ($N = 135$) collected between 16 February and 7 March 2023 (i.e., the same data set that was used as a basis for the power analysis reported back in Section 6.2.1). Each of these three prior specifications formally models a different kind of belief about future or yet-uninspected data regarding (a) what average additional delay on critical trials to expect in it (b) what effect magnitude to expect for the interaction of interest: With the uninformative prior specification, we basically pretend to be totally agnostic about whether critical trials would be processed faster or slower than control trials and also by what degree, although we are 95 % sure that it would be something within the log-milliseconds interval $[-0.98,\, 0.98]$, with our best guess being 0 log ms, i.e., no difference between critical and control trials. As for the interaction of interest between verification response and polarity, we would be equally agnostic, given the uninformative prior specification. That is, we would again find it equally plausible, *a priori*, for the sign of the effect to be either positive or negative, and our belief about its magnitude would likewise be that it is somewhere around 0 log ms, but with 95 % certainty not outside the boundaries of the interval $[-0.98,\, 0.98]$. Compare this to the meta-analysis-based prior specification where our best guess for the critical–control difference is 0.212 log ms (95 % credible interval: $[0.114,\, 0.310]$) and our best shot at guessing the effect magnitude of interest is -0.232 log ms (95 % credible interval: $[-0.422,\, -0.022]$). That is, when endorsing the meta-analysis posterior as our relevant prior, we are actually rather confident about seeing critical trials being processed more slowly than control trials and also quite optimistic about finding an interaction effect of the kind that the polarity hypothesis predicts ($\beta_{\texttt{VR:POL}} < 0$). Alternatively, if we were (perhaps most reasonably) to assume that the likelihood estimates for the two parameters in question that were obtained from (only) our own previous experimental data encode the most plausible expectation about what we might encounter in future data (yielded by the very same experimental paradigm!), then our belief would look like this: We would be even more confident to find a substantial positive critical–control difference (i.e., response delays on critical trials), with an intercept estimate of 0.349 $[0.255,\, 0.443]$ log ms, but rather undecided as to what sign to expect for the interaction effect of interest, with -0.093 $[-0.203,\, 0.017]$ log ms, although we would, in fact, be quite certain to find an effect magnitude that is not too far away from zero and somewhat more likely to be negative. A visualisation of the three different prior specifications for the magnitude of the `VR:POL` interaction effect of interest in form of a probability-density plot is provided in Figure 6.8.

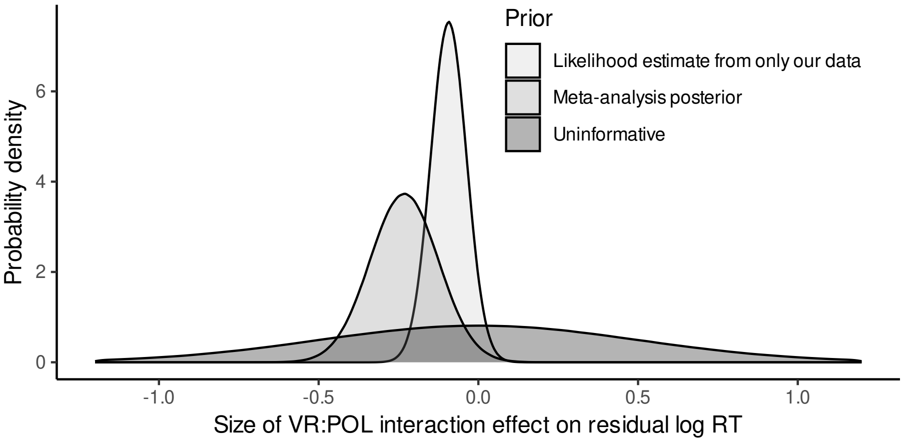


Figure 6.8: Probability densities of the three priors for $\beta_{\texttt{VR:POL}}$ specified in Table 6.3.

Let us move on to describing the Bayesian hierarchical linear model that we fit on Study Two's critical-trial response data, here, using the *brms* software package by Bürkner (2017). Its properties are summarised right below. The dependent measure, still residual log RT, remains the same as described earlier, i.e., is still constructed using the intermediate step of fitting the (frequentist) New Control Model on control trials (see Section 6.3) in order to residualise out the kind of variation caused by verification response, polarity, and their interaction that is not unique to critical trials. To be more precise, the description below actually accounts for three different models because—as we laid out—we fit the same type of model thrice, each time with a different prior specification for the intercept of residual log RT and the effect exerted on it by the `VR:POL` interaction. We may refer to this type of model as our Bayesian Model of Interest.

Bayesian Model of Interest (×3), on critical trials:

- Gaussian distribution with identity link function
- `res_log_RT` ~ `VR * POL + (VR * POL | subject) + (VR * POL | item)`
- Simplify group-level (i.e., random-)effects structure: For comparability, we adapt it to make it identical to the one of the eventually successfully converging frequentist counterpart, New Model (i), from Section 6.3.
- Population-level priors:
  - Intercept (α) and interaction slope ($\beta_{\texttt{VR:POL}}$): See Table 6.3.
  - Main-effect slopes: $\beta_{\texttt{VR}},\ \beta_{\texttt{POL}} \sim N(0,\ 0.5)$
  - Standard deviation: $\sigma \sim N_{+}(0,\ 1)$
- Group-level priors (ignore parameters not needed anyway after having simplified group-level effects' structure):
  - Standard deviations: $\tau_{s_1},\ \tau_{s_{\texttt{POL}}},\ \tau_{s_{\texttt{VR}}},\ \tau_{s_{\texttt{VR:POL}}},\ \tau_{i_1},\ \tau_{i_{\texttt{POL}}},\ \tau_{i_{\texttt{VR}}},\ \tau_{i_{\texttt{VR:POL}}} \sim N_{+}(0,\ 0.2)$

- By-subject correlations:
  $\rho s_{1,\,\texttt{POL}},\ \rho s_{1,\,\texttt{VR}},\ \rho s_{1,\,\texttt{VR:POL}},\ \rho s_{\texttt{POL},\,\texttt{VR}},\ \rho s_{\texttt{POL},\,\texttt{VR:POL}},\ \rho s_{\texttt{VR},\,\texttt{VR:POL}} \sim \text{LKJcorr}(2)$
- By-item correlations:
  $\rho i_{1,\,\texttt{POL}},\ \rho i_{1,\,\texttt{VR}},\ \rho i_{1,\,\texttt{VR:POL}},\ \rho i_{\texttt{POL},\,\texttt{VR}},\ \rho i_{\texttt{POL},\,\texttt{VR:POL}},\ \rho i_{\texttt{VR},\,\texttt{VR:POL}} \sim \text{LKJcorr}(2)$

A few quick explanations regarding notation: The model formula is given again in *brms/glmer* syntax. Where priors are defined, $N(\ldots,\ \ldots)$ denotes a normal distribution, while $N_+(\ldots,\ \ldots)$ stands for a *truncated* normal distribution of which only the positive, i.e., above-zero portion is considered. The subscripted s and i letters, used to disambiguate some group-level parameters' priors, are shorthand for subject and for item, respectively. The prior distribution chosen for any potentially included by-subject or by-item correlation terms is an LKJ correlation distribution with $\eta = 2$ (see Lewandowski et al., 2009), which is what the abbreviating notation LKJcorr(2) stands for.

For each of the three instances of the Bayesian Model of Interest, we compare the plausibility of its prior for $\beta_{\texttt{VR:POL}}$ given the data against that of a corresponding null model that assumes $\beta_{\texttt{VR:POL}} = 0$ (but still uses the same prior for α) by performing a Bayes factor analysis.

Results

The implementation of the three Bayesian Models of Interest and their corresponding null models which they are compared against has also been made available in the script **osf.io/ve9ab**; a reproduction of the results can be achieved by executing said script on the response data stored in **osf.io/y4pkm** until code line 575. All models were eventually fitted using the optimisation setting `adapt_delta = 0.99` and the hyperparameters `warmup = 2000`, `iter = 20000`, `cores = 4`, which resulted in successful fit without any divergence issues in all cases. As stated previously, the data the models were fitted on was the SPV response data from Study Two, collected on 6 April 2023. Since we had committed to applying the same group-level (a.k.a. random-)effects structure here as in our eventually converging, frequentist New Model (i) from Section 6.3, the present Bayesian models also employed the rather minimal structure that just includes by-subject and by-item group-level intercepts, but no group-level slopes whatsoever.

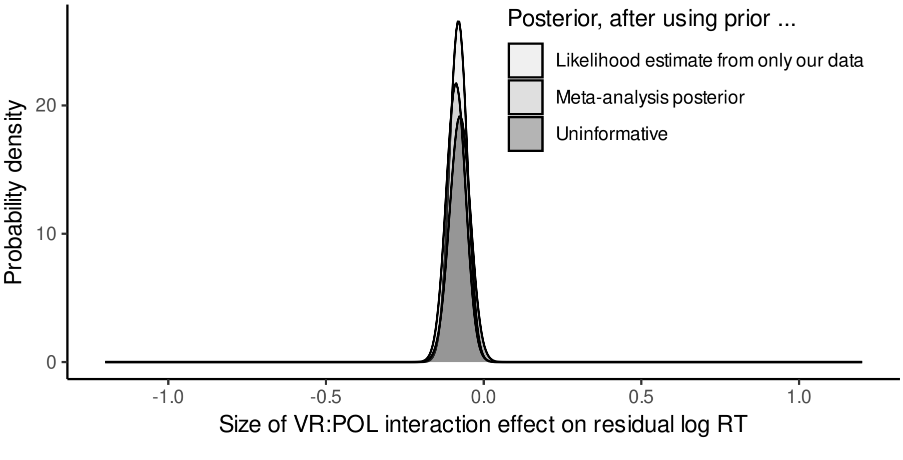


Figure 6.9: Probability densities of the three posteriors (normally approximated) for the interaction parameter of interest, $\beta_{\texttt{VR:POL}}$, after fitting variants of the Bayesian Model of Interest with three different prior specifications as given in Table 6.3.

Table 6.4: For each of three different prior specifications: The Bayes factor BF_{10} in favour of the Bayesian Model of Interest in comparison to a null model where the interaction parameter $\beta_{\texttt{VR:POL}}$ is constrained to zero.

Prior specification	Null	Alternative	BF_{10}
Uninformative	$\alpha \sim N(0, 0.5)$ $\beta_{\texttt{VR:POL}} = 0$	$\alpha \sim N(0, 0.5)$ $\beta_{\texttt{VR:POL}} \sim N(0, 0.5)$	0.84
Meta-analysis posterior	$\alpha \sim N(0.212, 0.050)$ $\beta_{\texttt{VR:POL}} = 0$	$\alpha \sim N(0.212, 0.050)$ $\beta_{\texttt{VR:POL}} \sim N(-0.232, 0.107)$	1.41
Lik. est. from only our data	$\alpha \sim N(0.349, 0.048)$ $\beta_{\texttt{VR:POL}} = 0$	$\alpha \sim N(0.349, 0.048)$ $\beta_{\texttt{VR:POL}} \sim N(-0.093, 0.056)$	6.56

For the three fitted Bayesian Models of Interest, the posterior estimates for the `VR:POL` interaction effect of interest are the following: When using the uninformative prior, we obtain the estimate $\hat{\beta}_{\texttt{VR:POL}} = -0.075$ $[-0.139,\ -0.010]$, on the log-milliseconds scale. When using the meta-analysis-based prior, the estimate is $\hat{\beta}_{\texttt{VR:POL}} = -0.088$ $[-0.150,\ -0.027]$. Lastly, when relying on the prior representing the likelihood estimate from only our own previous data, one ends up with a posterior estimate of $\hat{\beta}_{\texttt{VR:POL}} = -0.080$ $[-0.136,\ -0.024]$. Overall, it is apparent that these three estimates only marginally differ across each other, but that the estimate uncertainty (i.e., 95 % credible interval) is smaller the more informative the underlying prior has been. Further, the estimate yielded when using the uninformative prior specification ($\hat{\beta}_{\texttt{VR:POL}} = -0.075$) is the closest one to the frequentist likelihood estimate from our main analysis in Section 6.3.2 (also $\hat{\beta}_{\texttt{VR:POL}} = -0.075$, but which has been reported previously rounded to the second decimal point as -0.08). Figure 6.9 visualises the probability density of each of these posterior estimates for the effect of interest.

The Bayes factor BF_{10} represents the relative likelihood of a model of interest as compared to a null model of the same data. Its interpretation is rather straightforward in the sense that, e.g., a BF_{10} of 5 would signify that the model of interest is five times more likely than the null model. Typically, a Bayes factor greater than 3 is considered moderate evidence for the model of interest and one of greater than 10 strong evidence for said model. By contrast, a Bayes factor below $\frac{1}{3}$ is regarded as moderate evidence for the null model and a factor of smaller than $\frac{1}{10}$ as strong evidence for the null model. Thus, any Bayes factors between $\frac{1}{3}$ and 3 indicate that there is no decisive evidence in favour either of the model of interest or of the null model. For a concise discussion of the application of Bayes factors in the context of psycholinguistics, see Vasishth (2023).

As reported in Table 6.4, the variants of the Bayesian Model of Interest that use either the uninformative or the meta-analysis-based prior specification are associated with quite indecisive Bayes factors BF_{10} when compared against their respectively corresponding null models, i.e., a BF_{10} of 0.84 in case of the uninformative prior and a BF_{10} of 1.41 in case of the meta-analysis-based prior. Only the model variant that uses the prior derived from likelihood estimates from our own previous data successfully outperforms its corresponding null model in accounting for the newly collected data, with a BF_{10} of 6.56 (i.e., moderate evidence for the model of interest).

This result is interesting insofar as it adds some nuance to the binary conclusion about statistical significance obtained from our frequentist main analysis in Section 6.3: Seemingly, only if we adopt a strong prior belief about the magnitude of the effect of interest which is based on the data that we had collected earlier, we then eventually end up with an updated belief (once inspecting the new data) of such kind that it allows us to reject the null hypothesis in favour of the alternative with moderate confidence.

6.4.2 Re-Analysis à la Van Tiel and Pankratz (2021)

In the same vein as already done earlier with the data of Study One in Section 5.5.2, we want to carry out a re-analysis of the newly collected response data of Study Two using a frequentist paradigm that directly imitates the statistical method utilised by van Tiel and Pankratz (2021) in their study. The motivation for this endeavour is, again, to increase the comparability of our results to those reported by van Tiel and Pankratz.

As you may recall from what was laid out in Section 5.5.2, this method only considers trials featuring weaker scalemates (i.e., of trial types *t1…t3*±), but no trials featuring stronger scalemates (i.e., of trial types *t4…t6*±). Hence, we need to select a subset of the response data that exclusively encompasses critical-trial observations (i.e., *t1*±) together with those control-trial observations that are of type *t2*± or type *t3*± for the current statistical analysis.

The dependent measure, here, is simply log response time (`log_RT`). As predictors, condition (`COND`; i.e., critical vs. control), verification response (`VR`), and polarity (`POL`) are considered, but potential covariation due to response bias (`BIAS`; i.e., the scale-specific ratio of literal responses) or order of trial presentation (`ORD`) is also controlled for.

Here, binary variables are sum-coded, with critical condition, 'True' verification response, and positive polarity as > 0, and continuous variables are centred at a mean of 0 and re-scaled so as to show a standard deviation of 1.

Our first statistical model that is fitted here may be called the New VT&P'21-Like Omnibus Model. As fixed effects, it encompasses all of the predictors mentioned in the paragraph above as main-effect terms and additionally all possible two- or three-way interactions between condition, verification response, and polarity. Note that the interaction of interest for evaluating the polarity hypothesis is `COND:VR:POL`, in the present context. Further, the model includes by-subject and by-scale (here, *not* by-item) random intercepts, but no random slopes at all. Precisely following van Tiel and Pankratz (2021), it is fitted using some custom optimisation settings in *lme4* (Bates, Mächler, Bolker and Walker, 2015) which are given right below.

New VT&P'21-Like Omnibus Model, on all trials featuring a weaker scalemate:

- Gaussian distribution with identity link function
- `log_RT ~ COND * VR * POL + BIAS + ORD + (1 | subject) + (1 | scale)`
- Custom optimisation settings:
  - `optimizer = 'optimx', calc.derivs = FALSE, method = 'nlminb', starttests = FALSE, kkt = FALSE`

Again, on top of that, we also construct eight separate models for scale-specific subsets of the response data—this model type is labelled New VT&P'21-Like Scale-Specific Model. These models obviously do not include polarity as a predictor since it is a scale-specific property. In consequence, they are useful for telling whether a B&N effect is present, absent, or *reversed* (i.e., inversely present) for any particular examined scale, judging by the significance and the sign of the `COND:VR` interaction term.

New VT&P'21-Like Scale-Specific Model (×8), for each scale, on all scale-specific trials featuring a weaker scalemate:

- Gaussian distribution with identity link function
- `log_RT` $\sim$ `COND * VR + BIAS + ORD + (COND + VR || subject)`
- Custom optimisation settings:
  - `optimizer = 'optimx', calc.derivs = FALSE, method = 'nlminb', starttests = FALSE, kkt = FALSE`

Both the New VT&P'21-Like Omnibus Model and the eight different New VT&P'21-Like Scale-Specific Models are fitted with the restricted maximum likelihood (REML) criterion at convergence. For statistical tests on fixed-effect parameter estimates, the Satterthwaite (1946) procedure for degrees-of-freedom approximation is applied.

Results

The results reported here are reproducible by executing the script stored in **osf.io/ve9ab** on the data in **osf.io/y4pkm** up until line 677.

Estimates for the fixed-effect parameters from the fitted New VT&P'21-Like Omnibus Model are summarised in Table 6.5. We can see that the three-way interaction of interest does, indeed, display a significant effect ($\hat{\beta}_{\texttt{COND:VR:POL}} = -0.10$, $SE = 0.03$, $t = -2.86$, $p = .004$) with the expected sign ($\hat{\beta}_{\texttt{COND:VR:POL}} < 0$). Thereby, our main result from Section 6.3.2 is re-affirmed. In congruence with our previous results from this kind of analysis in Study One (Section 5.5.2), we also find significant main effects of condition ($\hat{\beta}_{\texttt{COND}} = 0.32$, $SE = 0.01$, $t = 39.42$, $p < .001$), of verification response ($\hat{\beta}_{\texttt{VR}} = -0.08$, $SE = 0.01$, $t = -10.57$, $p < .001$), and of order of presentation ($\hat{\beta}_{\texttt{ORD}} = -0.14$, $SE = 0.00$, $t = -35.71$, $p < .001$), hinting at the additional response slowdowns associated with critical trials, 'False' responses, and early presentation of trials, respectively.

Turning our attention to the eight fitted New VT&P'21-Like Scale-Specific Models, we can acknowledge that the `COND:VR` interaction is significant for the scales +possibility, +space, −time, and −space, but insignificant for the four remaining scales, as indicated by the overview in Table 6.6. Remarkably, the interaction patterns for +possibility (i.e., ⟨might be, is definitely⟩) and −time (i.e., ⟨did not always *VB*, never *VBD*⟩) constitute B&N effects (as $\hat{\beta}_{\texttt{COND:VR}} < 0$), while the interaction patterns for +space (i.e., ⟨somewhere *PP-LOC* it is, everywhere *PP-LOC* it is⟩) and −space (i.e., ⟨not everywhere *PP-LOC* is it, nowhere *PP-LOC* is it⟩) constitute *reversed* B&N effects (as $\hat{\beta}_{\texttt{COND:VR}} > 0$). This is a rather wild relationship considering that the polarity hypothesis ideally expects to find B&N effects consistently among positive scales, but reversed B&N effects consistently among (explicitly) negative scales. Instead, what we see here is one positive and one negative scale with a B&N effect, one positive and one negative scale with a reversed B&N effect, as well as two positive and two negative scales with no effect at all. At first glance, this is an even more puzzling result than what we saw earlier after performing the same kind of analysis on Study One's data. In fact, the only scale that behaved consistently across our two studies as well as in line with the polarity hypothesis' predictions is −space, showing a reversed B&N effect in both of our studies. All things considered, we can conclude that, despite finding an overall effect that is consistent with the polarity hypothesis (see our main result in Section 6.3.2), the predictions made by the polarity hypothesis fail to hold on a scale-specific level of analysis.

Table 6.5: Fixed-effect estimates from New VT&P'21-Like Omnibus Model, fitted to log RT, based on SPV response data collected on 6 April 2023 from 400 Prolific participants (post-exclusion). This table has a twin further above in Section 5.5.2 where analogous results from our previous study with smaller sample size are discussed: Table 5.5.

	$\hat{\alpha}$	*SE*	t	p
Intercept	7.56	0.04	194.98	< .001
	$\hat{\beta}$	*SE*	t	p
Condition (`COND`)	0.32	0.01	39.42	< .001
Verification response (`VR`)	−0.08	0.01	−10.57	< .001
Polarity (`POL`)	−0.18	0.07	−2.36	.064
Response bias (`BIAS`)	0.05	0.04	1.39	.222
Order of presentation (`ORD`)	−0.14	0.00	−35.71	< .001
Interaction of `COND:VR`	−0.04	0.02	−2.56	.011
Interaction of `COND:POL`	−0.07	0.02	−4.06	< .001
Interaction of `VR:POL`	−0.02	0.02	−1.30	.195
Interaction of `COND:VR:POL`	−0.10	0.03	−2.86	.004

Table 6.6: Estimates for the fixed effect of the interaction between condition and verification response, from each of the eight New VT&P'21-Like Scale-Specific Models, fitted to log RT, based on SPV response data collected on 6 April 2023 from 400 Prolific participants (post-exclusion). The rightmost column summarises if the effect of interest is either insignificant ('—') at $p < .05$, resembling a B&N effect ('present'), or resembling a *reversed* B&N effect ('reversed'). This table has a twin further above in Section 5.5.2 where analogous results from our previous study with smaller sample size are discussed: Table 5.6.

Scale	Interaction of `COND:VR`				
	$\hat{\beta}$	*SE*	t	p	B&N
+quantity	−0.04	0.05	−0.75	.455	—
+possibility	−0.34	0.06	−5.96	< .001	present
+time	0.04	0.05	0.98	.329	—
+space	0.19	0.06	3.05	.002	reversed
−quantity	0.02	0.05	0.45	.656	—
−possibility	0.01	0.05	0.12	.906	—
−time	−0.20	0.05	−4.21	< .001	present
−space	0.32	0.05	6.06	< .001	reversed

6.5 Unplanned Exploratory Analyses

In all of the statistical analyses reported so far, some potential sources of variance have been totally neglected because, implicitly, they were considered irrelevant. A salient example of this is how we have so far treated condition as a binary variable (critical/control), be it when we used this variable for residualisation or be it when we included it directly as a predictor in a Van-Tiel-and-Pankratz-inspired model. But recall that we are dealing with 12 different trial types, i.e., combinations of sentence variants $\in \{\exists, \forall, \neg\forall, \neg\exists\}$ and picture variants $\in \{0\,\%, 50\,\%, 100\,\%\}$, where it is by no means self-evident that any combinations which do not happen to be a critical trial type, i.e., neither [∃, 100 %] a.k.a. *t1+* nor [¬∀, 0 %] a.k.a. *t1−*, would all magically behave in the same way only because our theoretical and experimental requirements demand them to be treated as one holistic 'control condition'. Rather, it is quite plausible to suspect that there may also exist certain between-trial-type effect patterns worth accounting for. This intuition is somewhat supported by Figure 6.10 which plots mean raw response times (in milliseconds) grouped by trial type and by item: While a general dichotomy between critical (*t1*±) and control (*t2*...*t6*±) trial types does seem to roughly match the given data due to the markedly longer average response times on *t1*± trials, there are also discernible differences between, e.g., *t5*± and *t6*± trial types, especially when comparing the Possibility and Space items (see the 'X'-shaped interaction). But within the ontology of our previous statistical analyses, both *t5*± trials and *t6*± trials are all simply 'control trials' which, in this case, even expect the same unambiguously correct response ('False'). So, any response-behaviour differences that are sensitive to the *t5*±-versus-*t6*± distinction would be completely missed by all of our prior analyses.

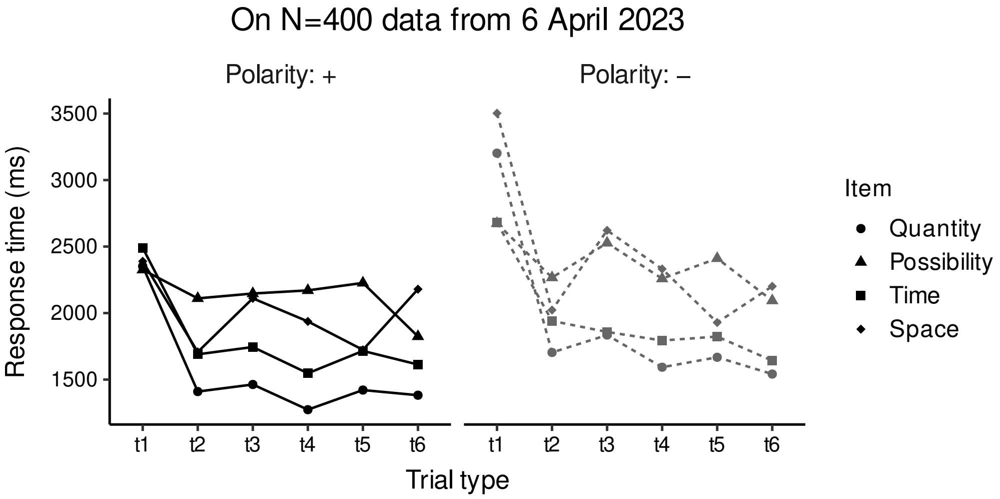


Figure 6.10: Mean raw response times (in milliseconds) as found in the SPV response data from 6 April 2023 that was collected from 400 Prolific participants (post-exclusion), grouped by trial type (*t1+*, *t2+*, ..., *t6+*, *t1−*, *t2−*, ..., *t6−*) and by item (Quantity, Possibility, Time, Space).

This is what motivates us, then, to conduct further exploratory analyses (decided upon spontaneously after data inspection) in which we try to model the response behaviour (i.e., response times/choices) shown by participants in a more systematic way, taking into account a broader range of possible predictor variables. Those very final analyses within the present work are reported right below in Section 6.5.1.

Table 6.7: Custom contrast-coding scheme for (within-polarity) trial type.

Trial-Type Contrast	***t1*±**	***t2*±**	***t3*±**	***t4*±**	***t5*±**	***t6*±**
Condition (critical vs. control)	0.834	−0.166	−0.166	−0.166	−0.166	−0.166
Expected VR (‘True’ vs. ‘False’)	0	0.6	−0.4	0.6	−0.4	−0.4
t2 vs. *t4*	0	0.5	0	−0.5	0	0
t3 vs. *t5*/*t6*	0	0	0.667	0	−0.333	−0.333
t5 vs. *t6*	0	0	0	0	0.5	−0.5

Table 6.8: Custom contrast-coding scheme for item.

Item Contrast	**Quantity**	**Possibility**	**Time**	**Space**
Quantity/Possibility vs. Time/Space	0.5	0.5	−0.5	−0.5
Quantity vs. Possibility	0.5	−0.5	0	0
Time vs. Space	0	0	0.5	−0.5

6.5.1 Comprehensive Models of Response Times and Choices

The two dependent measures we want to model here, separately, are log response time (`log_RT`) and verification response (`VR.bin`; i.e., ‘True’ = 1 or ‘False’ = 0).

As predictor variables for each, we consider the following: Since the previous analysis in Section 6.4.2, which imitated the method of van Tiel and Pankratz (2021), clearly demonstrated that the order of trial presentation (`ORD`) accounts for a substantial amount of variation in response times, we include it in the present analysis as well, while centring it at a mean of 0 and a standard deviation of 1. Further, polarity (`POL`) is included, as usual, as a sum-coded binary predictor, with positive polarity as >0. Then, crucially, we encode all six levels of the factor of within-polarity trial type (i.e., *t1*±, *t2*±, ..., *t6*±) as specified in Table 6.7—so that we do not only model differences for the contrast of critical vs. control condition (`COND`), but also for four further contrasts that meaningfully distinguish types of trials. In the previous analyses, we had always considered item (or scale) as a random effect. As we saw, however, resulting between-item differences in response times (and choices) have been so substantial that it would make sense to try modelling such between-item differences as fixed-effect predictors for once, considering that they may actually convey discrepancies that are not ‘random’, but rather reflecting something theoretically interesting that is grounded in their linguistic or visual properties. Therefore, we do, indeed, approach item as a four-level factor, here, and encode each of its levels using the custom contrast-coding scheme summarised in Table 6.8.

Beyond main effects, we are interested in modelling every two- or three-way interaction that can be derived by combining polarity, any trial-type-related contrast, and any item-related contrast. Note that, in total, this amounts to 38 different interaction terms.

We still want to model between-subject differences with by-subject random intercepts, but we do not consider any random slopes, here.

With everything stated above considered, we end up constructing two frequentist models, i.e., one *linear* mixed-effects model for log response time, and one *logistic* mixed-effects model for verification response. We can hereinafter refer to the former model as Comprehensive Model of Log Response Time and to the latter one as Comprehensive Model of Verification Response. Both models are implemented in *lme4* (Bates, Mächler, Bolker and Walker, 2015). Their formulas are given in (pseudo-)*glmer* syntax right below:

Comprehensive Model of Log Response Time, on the whole SPV data set:

- Gaussian distribution with identity link function
- `log_RT` ∼
  `ORD + POL * (COND +` [Expected VR] `+` [*t2* vs. *t4*] `+` [*t3* vs. *t5/t6*] `+` [*t5* vs. *t6*]`)`
  `* (`[Quantity/Possibility vs. Time/Space] `+` [Quantity vs. Possibility]
  `+` [Time vs. Space]`)`
  `+ (1 | subject)`

Comprehensive Model of Verification Response, on the whole SPV data set:

- Binomial distribution with logit link function
- `VR.bin` ∼
  `ORD + POL * (COND +` [Expected VR] `+` [*t2* vs. *t4*] `+` [*t3* vs. *t5/t6*] `+` [*t5* vs. *t6*]`)`
  `* (`[Quantity/Possibility vs. Time/Space] `+` [Quantity vs. Possibility]
  `+` [Time vs. Space]`)`
  `+ (1 | subject)`

Once fitting these two models, statistical significance for all fixed-effect parameter estimates is computed either based on the Satterthwaite (1946) procedure or on the Wald (1943) z-test, depending on whether the model in question is linear or logistic. Because of the huge number of statistical comparisons to be performed here, we restrain ourselves to only considering $p < .001$ cases as potentially relevant.

Results

The models described above were successfully fitted on the $N = 400$ SPV response data from 6 April 2023 without any convergence issues.[25] This process can be reproduced by running the script from **osf.io/ve9ab** on the data stored in **osf.io/y4pkm** until the very end, i.e., line 810 of the R code.

As for the Comprehensive Model of Log Response Time, all of its fixed-effect estimates that are significant at a threshold of $p < .001$ are listed in Table 6.9. Analogously, all significant ($p < .001$) fixed-effect estimates stemming from the Comprehensive Model of Verification Response are reported in Table 6.10.

To discuss each and every one of these exploratorily detected significant effects would lead us a bit to far, so we are going to pick out only a selected few that are interesting to think about.

Since we have already observed in earlier analyses that there are very salient presentation-order effects on response times, detecting this kind of effect here again should come as no surprise ($\hat{\beta}_{\texttt{ORD}} = -0.148$, $SE = 0.002$, $\boldsymbol{t} = -60.43$). However, quite remarkably, we also find a presentation-order effect on verification response itself ($\hat{\beta}_{\texttt{ORD}} = -0.161$, $SE = 0.028$, $\boldsymbol{z} = -5.77$), suggesting that people become more and more likely to give ‘False’ rather than ‘True’ responses as trials go by. There is also a significant main effect of polarity both on log response time ($\hat{\beta}_{\texttt{POL}} = -0.141$, $SE = 0.005$, $\boldsymbol{t} = -28.75$) and on verification response ($\hat{\beta}_{\texttt{POL}} = 0.416$, $SE = 0.094$, $\boldsymbol{z} = 4.43$). It indicates that positive sentences are verified faster than negative sentences and that, generally speaking, positive sentences provoke a ‘True’ response in participants more often than negative sentences do. Moving on to trial-type-related contrasts, we find that control trials that demand a correct

[25] But note that the Comprehensive Model of Verification Response was generously fitted with the custom optimisation setting `maxfun = 500000` in order to ensure successful convergence before the maximum number of iterations would be reached.

'True' response are verified a bit faster than other control trials which demand a correct 'False' response instead ($\hat{\beta}_{\text{Expected VR}} = -0.020$, $SE = 0.005$, $\boldsymbol{t} = -3.67$). But, in addition, we also see substantial differences in response times within the larger group of 'False'-demanding control trials: Those trials among them that belong to the *t3*± trial type are verified more slowly than those belonging to type *t5*± or *t6*± ($\hat{\beta}_{t3\,\text{vs.}\,t5/t6} = 0.097$, $SE = 0.007$, $\boldsymbol{t} = 13.27$). In turn, *t5*± trials are still responded to more slowly than *t6*± trials ($\hat{\beta}_{t5\,\text{vs.}\,t6} = 0.036$, $SE = 0.008$, $\boldsymbol{t} = 4.21$), on average. Regarding items, our models detect, e.g., a salient discrepancy between the Quantity and the Possibility item that affects both log response time ($\hat{\beta}_{\text{Quantity vs. Possibility}} = -0.284$, $SE = 0.007$, $\boldsymbol{t} = -41.03$) and verification response ($\hat{\beta}_{\text{Quantity vs. Possibility}} = -0.543$, $SE = 0.123$, $\boldsymbol{z} = -4.40$). We can interpret this as suggesting that trials of the Quantity item are generally processed faster than trials of the Possibility item, but also that people are rather inclined to judge trials of the Quantity item to be 'False' than they are to do so with trials of the Possibility item. Moreover, there are also quite a few interaction terms that come out as significant, here. One that is useful to illustrate is the two-way interaction $[t5\,\text{vs.}\,t6] : [\text{Time vs. Space}]$ ($\hat{\beta}_{[t5\,\text{vs.}\,t6]\,:\,[\text{Time vs. Space}]} = 0.245$, $SE = 0.024$, $\boldsymbol{t} = 10.25$) because its effect pattern can be easily acknowledged visually above in Figure 6.10 (see the '<'-shaped interaction). Simultaneously, this interaction term affects verification response as well ($\hat{\beta}_{[t5\,\text{vs.}\,t6]\,:\,[\text{Time vs. Space}]} = 1.705$, $SE = 0.435$, $\boldsymbol{z} = 3.92$). Finally, among three-way interactions, the term [Polarity] : [Condition] : [Time vs. Space] is worth mentioning as it also affects both response time ($\hat{\beta}_{[\text{Polarity}]\,:\,[\text{Condition}]\,:\,[\text{Time vs. Space}]} = 0.213$, $SE = 0.037$, $\boldsymbol{t} = 5.73$) and response choice ($\hat{\beta}_{[\text{Polarity}]\,:\,[\text{Condition}]\,:\,[\text{Time vs. Space}]} = -4.037$, $SE = 0.412$, $\boldsymbol{z} = -9.80$), but an attempt at interpretation of the complex interplay of factors suggested by it lies outside the current scope of our interest.

Overall, it must be recognised from this exploratory analysis that there is much more going on which affects our dependent measures of interest than all our previous analyses could possibly have detected. Although we are not allowed to conclude anything theoretically meaningful from any of the effects reported here just yet, due to the exploratory nature of this endeavour, many of them could surely inspire future, confirmatory work to try and replicate certain effects if a plausible theoretical explanation for them can be thought of in advance.

Table 6.9: All fixed-effect estimates from the Comprehensive Model of Log Response Time, fitted on SPV response data collected on 6 April 2023 from 400 Prolific participants (post-exclusion), where significance of $p < .001$ is reached.

Effects on Log Response Time ($p < .001$)			
	$\hat{\alpha}$	*SE*	t
Intercept	7.483	0.012	623.71
	$\hat{\beta}$	*SE*	t
Order of presentation (`ORD`)	−0.148	0.002	−60.43
Polarity (`POL`)	−0.141	0.005	−28.75
Trial Type			
– Condition (`COND`; critical vs. control)	0.344	0.007	52.38
– Expected VR (‘True’ vs. ‘False’)	−0.020	0.005	−3.67
– *t3* vs. *t5/t6*	0.097	0.007	13.27
– *t5* vs. *t6*	0.036	0.008	4.21
Item			
– Quantity/Possibility vs. Time/Space	−0.047	0.005	−9.68
– Quantity vs. Possibility	−0.284	0.007	−41.03
– Time vs. Space	−0.140	0.007	−20.18
Two-Way Interactions			
– [Polarity] : [Condition]	−0.094	0.013	−7.17
– [Polarity] : [*t3* vs. *t5/t6*]	−0.063	0.015	−4.33
– [Polarity] : [Quantity vs. Possibility]	−0.097	0.014	−7.05
– [Polarity] : [Time vs. Space]	0.095	0.014	6.89
– [Condition] : [Quantity vs. Possibility]	0.411	0.019	22.10
– [Expected VR] : [Time vs. Space]	0.073	0.015	4.69
– [*t2* vs. *t4*] : [Quantity/Possibility vs. Time/Space]	0.068	0.017	4.04
– [*t2* vs. *t4*] : [Time vs. Space]	0.187	0.024	7.81
– [*t3* vs. *t5/t6*] vs. [Time vs. Space]	−0.079	0.021	−3.81
– [*t5* vs. *t6*] : [Quantity/Possibility vs. Time/Space]	0.141	0.017	8.32
– [*t5* vs. *t6*] : [Quantity vs. Possibility]	−0.099	0.024	−4.15
– [*t5* vs. *t6*] : [Time vs. Space]	0.245	0.024	10.25
Three-Way Interactions			
– [Polarity] : [Condition] : [Time vs. Space]	0.213	0.037	5.73
– [Polarity] : [Expected VR] : [Quantity/Possibility vs. Time/Space]	0.086	0.022	3.96

Table 6.10: All fixed-effect estimates from the Comprehensive Model of Verification Response, fitted on SPV response data collected on 6 April 2023 from 400 Prolific participants (post-exclusion), where significance of $p < .001$ is reached.

Effects on Verification Response ($p < .001$)			
	$\hat{\alpha}$	*SE*	z
Intercept	−0.490	0.054	−9.02
	$\hat{\beta}$	*SE*	z
Order of presentation (`ORD`)	−0.161	0.028	−5.77
Polarity (`POL`)	0.416	0.094	4.43
Trial Type			
– Condition (`COND`; critical vs. control)	1.134	0.070	16.30
– Expected VR (‘True’ vs. ‘False’)	6.947	0.124	55.87
Item			
– Quantity vs. Possibility	−0.543	0.123	−4.40
Two-Way Interactions			
– [Polarity] : [Expected VR]	1.727	0.238	7.25
– [Polarity] : [Quantity/Possibility vs. Time/Space]	−0.857	0.188	−4.57
– [Condition] : [Quantity vs. Possibility]	−0.947	0.185	−5.11
– [Condition] : [Time vs. Space]	−1.271	0.206	−6.17
– [Expected VR] : [Quantity vs. Possibility]	2.618	0.293	8.93
– [Expected VR] : [Time vs. Space]	1.667	0.375	4.44
– [*t2* vs. *t4*] : [Time vs. Space]	−2.289	0.672	−3.41
– [*t5* vs. *t6*] : [Time vs. Space]	1.705	0.435	3.92
Three-Way Interactions			
– [Polarity] : [Condition] : [Quantity vs. Possibility]	−1.227	0.370	−3.32
– [Polarity] : [Condition] : [Time vs. Space]	−4.037	0.412	−9.80
– [Polarity] : [Expected VR] : [Quantity vs. Possibility]	2.276	0.586	3.88

Chapter 7
Discussion

Where does all this leave us now? We started out with a hypothesis about scalar implicatures (SIs) and polarity that was initially brought forth by van Tiel et al. (2019) and van Tiel and Pankratz (2021), and we designed a crowd-sourcing experiment to test that hypothesis. Eventually, this resulted in us conducting two experimental studies, the first one with a post-exclusion sample size of $N = 100$ and the second one with $N = 400$. On top of that, we were inspired by earlier work on individual differences in pragmatic processing to assess individual cognitive properties of participants of our first study along the dimensions of working memory capacity, print exposure, and fluid intelligence. Then, we exploratorily tried to relate differences along these dimensions to response behaviour in our main, SI-processing-related experimental task (sentence–picture verification). What can we take away from the obtained results reported in the previous sections? Have we discovered something? Was it all worth it? Let us find out.

7.1 Evidence Regarding Polarity Hypothesis

Our first experimental study (Study One; Chapter 5) did not show the effect predicted by the polarity hypothesis that van Tiel and colleagues have proposed. However, we have argued that this null result should not eagerly be viewed as *evidence against* the polarity hypothesis, considering that our sample size in Study One ($N = 100$) might have been too small to be able to detect the desired effect. Mainly by this consideration, we were then motivated to carry out our follow-up experimental study (Study Two; Chapter 6) whose sample size was considerably larger and had been decided upon based on a prospective power analysis. As it turned out, the data collected from the experiment of Study Two ($N = 400$) did, in fact, show the predicted interaction between verification response and polarity on response times to be significant, although with a rather small estimated effect magnitude of 0.08 log ms (i.e., here, about 135 ms). Recall that an attempt at synthesis of all prior evidence in form of a meta-analysis had suggested that a magnitude of about 0.23 log ms should be expected. Thus, we can say to have obtained rather mild support for the polarity hypothesis. Nevertheless, some concerns about the generalisability even of this mildly favourable 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 Horn scales), we find that only 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. This observation is further underlined by a particular exploratory analysis that was carried out next to our main, hypothesis-driven one: In an attempt to statistically re-analyse our collected response data following an approach as close as possible to the one employed by van Tiel and Pankratz (2021), for comparability purposes, we showed that when evaluating the polarity hypothesis' claims on a scale-by-scale basis, the fact whether a scale would display a B&N effect (i.e., pragmatic responses take longer) or a reversed B&N effect (i.e., literal responses take longer) was not satisfactorily accounted for by the polarity hypothesis at all in the collected experimental data. Furthermore, the results of a conducted Bayes factor analysis of the same data add even more nuance to the basic conclusion of having found (mild) evidence in favour of the polarity hypothesis: Only under the adoption of a highly informative prior centred around an effect-size estimate drawn from our own previously collected data (i.e., from a pilot study and Study One), the resulting Bayes factor becomes large enough (> 3) to indicate some degree of support for the polarity hypothesis.

In summary, the present results may provide a useful source of evidence (among others) regarding SI processing in general and the polarity hypothesis in particular, which future work is invited to take into account and build upon. However, they are not exactly suitable for yielding a decisive, yes-or-no answer to the question whether the polarity hypothesis by van Tiel and colleagues, including its theoretical justification grounded in the difficulty of negation processing, is true or false after all. Given that its predicted—and, to some extent, detected—effect pattern does not even generalise consistently across our very narrowly selected linguistic materials that were tested here, it is even less obvious why someone should expect it to generalise more broadly to any within-English cases of scalar implicature, let alone crosslinguistically. In fact, the issue of a potential lack of generalisability ties the present results in quite well with a recently emerged strain of research on SIs that is concerned with *scalar diversity* (see, e.g., Doran et al., 2009; van Tiel et al., 2016), i.e., describing and coming up with ways to account for the salient between-scale variability that is observed in how people process SIs derived from a broad range of different Horn scales. Ironically enough, the polarity hypothesis itself was born out of that context of research as a theoretical model of yet another property of Horn scales (i.e., polarity) that can explain why people treat SIs from certain scales differently than SIs from other scales. But given our present results, it seems that even the very effect pattern predicted by the polarity hypothesis may really itself be a victim to the all-encompassing phenomenon of scalar diversity since it manifests itself very differently (if at all) across scales. A potential and valid objection to this idea could be, of course, that it is precisely other, confounding properties of scales, aside from polarity, that explain the large between-scale variability we observe in our data with regard to the interaction of interest on RTs between polarity and response. We return to a discussion of this point below in Section 7.3. Another potential objection could be that our employed experimental method has particular flaws that prevent it from consistently detecting the predicted effect pattern on a scale-by-scale basis. These kinds of objections are discussed in Section 7.4.

Finally, recall that, in Study One, we actually did not only test *the* polarity hypothesis (i.e., concerning differences in response times), but also a secondary, more speculative hypothesis about a modulation of frequencies of literal vs. pragmatic responses directly by a main effect of polarity. As we obtained a relatively clear null result with regard to that secondary hypothesis in our first experimental study and since this hypothesis is not extensively backed up by prior work anyway (but see Gotzner et al., 2018), we did not attempt to test it again in our self-replicatory experiment of Study Two.

7.2 What About Individual Differences?

The question of individual differences (IDs) was addressed exclusively in Study One ($N = 100$). In particular, it was examined if accounting for participant-specific estimates of working memory capacity (WMC), fluid intelligence, or print exposure would aid in predicting response behaviour during SI processing beyond what is modelled already by the predictor variables relevant to the polarity hypothesis.

Recall that each cognitive construct of interest was assessed using one single task: WMC using an operation span (OSpan) task, fluid intelligence using Raven's Progressive Matrices (RPM), and print exposure using an author recognition test (ART). After performing multiple processes of backwards model selection, the only two ID-based predictors that were found to significantly improve model fit to the data were a salient main effect of fluid intelligence and a moderate three-way interaction between print exposure, verification response, and polarity. The effect of fluid intelligence was such that it showed more intelligent individuals to generally slow down more drastically in the face of SI-induced pragmatic ambiguity than less intelligent individuals. The directionality of the three-way interaction involving print exposure, although rather likely to be the result of a statistical fluke, is intuitively accessible in the sense that the two-way effect pattern predicted by the polarity hypothesis is more and more accurately approximated the less an individual has been exposed to printed language in the past.

As had been summarised early on in Chapter 2, there were various reasons as to why to suspect to see an interplay between IDs in WMC, fluid intelligence, or print exposure, on the one hand, and polarity-contrastive SI processing behaviour, on the other hand. Do these *a priori* reasons accurately reflect the two potential patterns of effects exploratorily stumbled upon here?

A finding from earlier work that rather closely falls in line with our present one regarding the main effect of fluid intelligence had been reported by Ryzhova et al. (2023): There, fluid intelligence (also assessed through an RPM task) was found to positively correlate with how likely people would be to draw pragmatic inferences in a task involving the comprehension of informationally redundant utterances. Note that the dependent measure that the main effect of intelligence acts upon is a different one there (i.e., whether a response is literal or pragmatic) than it is here (i.e., response times). So, our results would have been more consistent with the Ryzhova et al. finding if we had also or instead detected a main effect of intelligence on verification response itself—which was tried, but did not show significance. Nonetheless, under rather liberal interpretative assumptions, one could say that both findings (ours and the Ryzhova et al. one) point towards the idea that fluidly more intelligent human comprehenders are cognitively more sensitive to pragmatic ambiguity. On the contrary, none of the reviewed findings in earlier work concerning print exposure—be it in the context of pragmatic or negation processing—reflect the directionality of the three-way effect pattern obtained here, hence raising the suspicion even further that it is simply statistical noise.

In a separate follow-up analysis, we examined the distribution of responder types, i.e., the distribution of subjects grouped by the degree to which they were inclined to provide literal or, rather, pragmatic responses in the SPV task. The take-away was that there was no strong tendency towards extreme consistency on either side of the spectrum, but that fully consistent literal responders were greater in number than fully consistent pragmatic responders. Based on a responder-type grouping using arbitrary cut-off values, we then focused on the subgroup of literal-leaning responders in particular. Back in Chapter 2, we had discussed a rather specific prediction about that subgroup, stating that there may be differences in how and why someone ends up being a literal-leaning responder, with the expectation that more print-exposed or more intelligent people would often draw an SI, but consciously decide to ignore it and thus respond literally (longer RTs), whereas less print-exposed or less intelligent people would mostly not draw an SI in the first place and therefore straightforwardly respond literally (shorter RTs). Regarding fluid intelligence, but not print exposure, this prediction was, indeed, confirmed by a statistical analysis of only the literal responses given by the subgroup of responders in question. However, given our knowledge of the overall main effect of fluid intelligence on RTs, which is not constrained to this subset of data in particular, one should better avoid viewing the confirmation of said prediction about literal responders as really demonstrating something that is specific to only those responders.

One issue that remains to be discussed regarding the assessments of individual cognitive properties made here is the so-called task impurity problem (see, e.g., Schweizer, 2007). That is, assuming that the cognitive constructs WMC, fluid intelligence, and print exposure are psychologically real, it is not at all obvious—in fact, it is unlikely—that each of them can be perfectly measured using one single experimental task: Consider that any single task has specific idiosyncrasies in its design that inevitably introduce additional variation in the dependent measure due to benign reasons and hence not due to whatever latent cognitive construct is of interest. The discussion of this problem is also followed up upon further below, in Section 7.4.

7.3 Conceptual Issues

In this section, we want to address some concerns one might have about the theoretical and conceptual underpinnings of the research questions investigated here. Then, right below, in Section 7.4, we are going to discuss to what extent the methods utilised here have been suitable for tackling these research questions.

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 Horn 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 Horn 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. Secondly, given any particular Horn 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, where $1 \leq i < k \leq n$. 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 $n \neq 2$), 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. 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., appears to confirm it.

Now, as was hinted at in Section 7.1, it would be useful to consider if there are differences in properties like availability or distinctness between the Horn scales that we ourselves have examined in our presently reported studies. If this is the case, then potential objections against our claim to have found the polarity hypothesis not to generalise across scales should be taken more seriously; it may then very well be that such differences in other properties than polarity are confounding the observed effect patterns. Recall the eight Horn scales that were examined by us:

- +quantity: ⟨some, all⟩
- −quantity: ⟨not all, none⟩
- +possibility: ⟨might be, is definitely⟩
- −possibility: ⟨might not be, is definitely not⟩
- +time: ⟨sometimes *VBD*, always *VBD*⟩
- −time: ⟨did not always *VB*, never *VBD*⟩
- +space: ⟨somewhere *PP-LOC* it is, everywhere *PP-LOC* it is⟩
- −space: ⟨not everywhere *PP-LOC* is it, nowhere *PP-LOC* is it⟩

Let us look at the property of distinctness first. We would argue that all of these eight scales are bounded since the strongest term of each does, in fact, represent a logical end point on its respectively spanned meaning dimension. Further, as we have seen in Chapter 4, all of the scales display a logical norm of correctness (Dieussaert et al., 2011) as they resemble a well-defined quantification pattern of either $\langle \exists, \forall \rangle$ or $\langle \neg \forall, \neg \exists \rangle$ 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). As van Tiel et al. (2016) theorise that availability is higher in case of closed grammatical classes (say, pronouns) than it is with open grammatical classes (say, adverbs), 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 Horn scales: For instance, a Google search of `"some of"` yields 4.3 B hits and a search of `"all of"` 3.9 B hits, implying a relative frequency of the weaker scalar 'some' with respect to the stronger one 'all' of about 1.1. But a Google search of `"might be"` yields 2.8 B hits and a search of `"is definitely"` only 0.2 B hits, thus pointing towards a relative frequency of 'might be' with respect to 'is definitely' of about 14. Since $1.1 \not\approx 14$, 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 in comparison to the stronger one yielding SIs less often than scales where the opposite is the case, according to the reasoning by van Tiel et al. Similar arguments can be probably be made with regard to the third subproperty of availability alluded to by van Tiel et al., i.e., 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 Horn 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. Nonetheless, it is still quite dubious how such an availability-caused modulation of effect patterns (here, on response times) would look like, given that van Tiel et al. (2016) themselves did not even obtain significant results in support of any of the ways in which they had hypothesised scale availability to impact SI processing behaviour.

Aside from the above issues related to violations of the often-held uniformity assumption regarding SIs, there is also another traditional assumption about SIs that merits to be critically questioned: the *homogeneity assumption.* Rather than being concerned with supposing a lack of between-scale variability (like the uniformity assumption), the homogeneity assumption supposes that there is no *within-scale* variability. That is, any SIs derived from the same particular Horn scale are assumed to all behave in the same way, unaffected by differences in their linguistic contexts. Although, already intuitively, this alleged context-independence of SIs seems kind of hard to defend, it is actually a crucial 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 a 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 and 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 in form of our employed sentences and pictures. This concern directly points towards a methodological problem which we are going to elaborate on below in Section 7.4.

Moreover, it has recently also been called into question if it is cognitively plausible that people compute hypothetical, stronger alternatives to a weak scalar in form of concrete linguistic strings, in their minds. That is, a tacit assumption made by the entire endeavour of conceptualising SIs on the basis of Horn scales, i.e., ordered lists of concrete linguistic strings, is that it is precisely these strings which are cognitively accessed (say, from the mental lexicon) when deriving an SI. Rather, an alternative view (Buccola et al., 2022; Hu et al., 2022) holds that hypothetical alternatives are cognitively represented as abstract semantic concepts which may, but do not have to correspond to only one particular linguistic form. In designing our present experiments, we did not pay much attention to this idea. But judging just from common sense, it does seem quite relevant to address the idea especially in regards to our +possibility and −possibility scales, i.e., ⟨might be, is definitely⟩ and ⟨might not be, is definitely not⟩: It would be foolish to assume that a listener or reader of a sentence with 'might be' would literally compute a hypothetical alternative with exactly the string 'is definitely' in their mind only to then be able to reject the truth of said alternative out of pragmatic considerations. Instead, intuition tells us that such an alternative could also be represented through various other linguistic forms like 'must be', 'has to be', 'is certainly' etc.—if it is grounded in any particular linguistic form at all. But then the explicit exposure to our intended stronger alternatives with the string 'is definitely (not)' in some of the administered *control* trials (i.e., of trial types *t4*±, *t5*±, *t6*±), may have primed or corrupted some participants' natural sense of mentally representing alternatives to 'might (not) be' and could thus have modulated their recorded response behaviour on *critical* trials in unforeseen ways.

Going even a step further, a very recent and provocative study by Kissine and De Brabanter (2023) claims that SI derivations due to underinformative utterances are neither string-based nor concept-based, but rather do not exist at all. That is, they do not exist in the commonly assumed sense that comprehenders would cognitively compute and then negate a hypothetical, stronger alternative to the surface utterance. Kissine and De Brabanter argue for this novel viewpoint on the basis of experimental results that show participants to prefer to retro- and introspectively justify their choice of a pragmatic response without recurring to the idea of a negated alternative (like 'but not all' for the scalar 'some'), even while a justification based on this idea is offered to them as one of two possible justification options to select from.

If the view held by Kissine and De Brabanter does indeed reflect (cognitive) reality, then this would drastically shake the foundations of many of the theoretical arguments that have been made around scalar implicature processing—including, in fact, those for the polarity hypothesis. Remember that van Tiel and Pankratz (2021)'s main theoretical argument for embracing the polarity hypothesis involved linking the fact that negative information takes long to process with the idea that a mentally computed alternative to a surface utterance either would or would not contain negative information, depending on the polarity of the scalar in that surface utterance. But this entire argument falls apart in case that there is no such thing as a computed alternative to begin with.

Lastly, our provided definition of a Horn scale in Chapter 2 may be seen by some as too lenient in the sense that it allows for linguistic strings of arbitrary type and length to be considered scalars as long as they can be replaced with an alternative string in such a way that the resulting sentence variants can be meaningfully ordered by entailment. We suspect that a scalar like 'might (not) be' or like 'is definitely (not)' would be viewed by some rather syntactically focused pragmaticists as ill-formed since it is not a constituent (i.e., a leaf or a full branch of a generative syntactic tree). In fact, typical examples of Horn scales found in previous literature consist of scalars that are just single words—hence why the quasi-synonymous term *lexical scale* is often used to refer to a Horn scale. All that said, our present work is mainly grounded in the theoretical discussions encountered in the papers by van Tiel et al. (2019) and van Tiel and Pankratz (2021); and in those, at least the scale ⟨not all, none⟩ is sometimes mentioned. In a framework where the linguistic string 'not all' can be viewed as a scalar, while neither being a single word nor, at least, a constituent, we see no reason why one should not be allowed to grant the status of a scalar to a more complex string like 'is definitely not' (or eventually even 'did not always hit') as well.

7.4 Methodological Issues

As already briefly brought up by the end of Section 7.1, our present experimental studies may suffer from a task impurity problem (Schweizer, 2007), meaning that they rely on operationalising every cognitive construct of interest (be that SI processing or WMC / fluid intelligence / print exposure) using a single behavioural task (be that SPV or OSpan / RPM / ART). This is not ideal because any single task's design will necessarily give rise to patterns of variation in the response measure that are really just noise and have nothing to do with the cognitive construct of interest—unless of course that cognitive construct actually happens to be 'OSpan performance', 'ART performance', or the likes.

For example, consider the author recognition test (ART). With it, we are really interested in assessing the degree to which an individual has been exposed to printed language in the past, so that we would hope that this degree is accurately reflected in the ART performance scores that we obtain. But unfortunately, there are several confounding factors that may be at play in shaping the obtained performance scores: First, as the ART score incorporates an additional penalty for any fake author name falsely classified as real—hence encouraging the rational participant to focus on precision rather than recall—we may end up seeing two participants A and B who, in reality, are print-exposed to the same degree, but where A has a lower score than B due to being someone who enjoys taking irrational risks, e.g., making bold guesses about fake author names being real in the face of uncertainty. Second, there may be two participants C and D who are also equally print-exposed, but where C has a lower score than D due to having an unusual preference for exclusively reading niche folklore fairy tales and instruction manuals, thus disfacilitating the successful recognition of rather mainstream real author names. Third, with our particular implementation of ART, responses to any trial would be given by pressing either the key **1** (= recognised as real author) or **0** (= not recognised as real author). On conventional keyboards, the **1** key is located to the very left, but the **0** key to the very right. If, then, there were two equally print-exposed participants, a left-handed person E and a right-handed person F, then we may see E naturally pressing the **1** key more often than her fellow F simply because she is more at ease, in case of a trial that she is doubtful about, to use her dominant hand to respond whatever. This might lead E to end up with a lower score than F due to unintentionally collecting overall more penalty points because of false-positive classifications with the **1** key. When administering the ART task, we typically do not want to measure personality traits, print-genre preferences, or handedness. Nonetheless, the impurification of our measure of interest by such confounds is almost impossible to avoid.

This was just to illustrate, of course, how the task impurity problem can affect virtually any kind of experimental task; it is not an issue tied to the ART task in particular by any means. So what can be done, then, to mitigate this problem? One solution is to administer multiple different tasks that tap into the same construct of interest and, then, to calculate a latent variable which consists only of the variance shared across manifest measures from individual tasks (as done, e.g., by Miyake et al., 2000, with regard to executive functions). To take the example of ART again, a solution of this kind might lead to an experimental design where ART is administered alongside a self-report questionnaire about reading habits in an attempt to compute one latent variable of print exposure from the combined responses to these tasks. In any case, it also has to be acknowledged that the task impurity problem is often treated as an issue of mere theoretical, but not real practical concern. That is because, in most cases when an experimental study in psycholinguistics has to be designed, it would simply not be feasible to go for a fully fledged latent variable analysis for each and every cognitive construct that one wants to assess—due to budget, sample-size, and time constraints.

A somewhat related methodological problem concerns the issue that the sentence–picture items we employed in our two experiments may not necessarily be representative of how the polarity-contrastive pairs of Horn scales they instantiate behave in a broader range of (real-world) conversational contexts—that is, we have an 'item impurity' problem. Here, this issue is further aggravated by the fact that we only used one (type of) sentence and one (type of) picture per pair of scales. In consequence, it is hard to rule out that any extraordinary effects observed for, say, the Space item and thus eagerly attributed to properties of the Horn scales +space and −space are not really just effects of processing a named entity instead—i.e., the word 'Africa' in the sentence 'Somewhere in Africa it is daytime.' or one of its variants. Likewise, we cannot be sure if some observed differences between how the Time item and how the Quantity item is responded to, on average, are really linked to differences concerning scalars like 'sometimes $_{VBD}$' versus scalars like 'some' in general, or if it rather has to do with it being harder to visually discern the location of thin arrows on a cartoon-like archery target than to judge the colour(s) of a visually very salient set of 16 identical-looking drawings of apples. In a future study looking into the same Horn scales, it may thus help to design an experiment that instantiates each scale within several different visual/linguistic contexts rather than just a single one.

In Chapter 4, while talking about the conducted pilot studies, we mentioned that one crucial aspect in which our very first two pilot studies (from 28 January and 13 February 2023) had differed from the last pilot study (16 February 2023) and eventually from the two actual studies (6 & 7 March 2023 and 6 April 2023) was how the verification task was framed to participants. Whereas, initially, we had opted for an instructional framing that would explicitly ask the participant for a truth-value judgement, i.e., if a sentence was 'True' or 'False' with respect to a picture, we eventually changed this to an alternative framing where the participant would rather be asked about whether a sentence was 'a good description' or 'a bad description' of the picture following it—hence actually asking for a judgement on *felicity* rather than truth instead. Our decision for this change in framing was motivated mainly by practical reasons: In the pilot studies, we observed that the truth-based framing would yield much fewer pragmatic responses (between 10.0 % and 20.8 %) than the felicity-based framing (38.8 %). Further, the studies by van Tiel et al. (2019) and van Tiel and Pankratz (2021), in whose tradition we wanted to operate, also employed a felicity-based framing in their administered SPV tasks. Looking at prior literature more broadly, it generally seems that studies administering a *sentence–picture* verification task prefer to use a felicity-based paradigm (Tavano and

Kaiser, 2010; van Tiel et al., 2019; Marty et al., 2020; van Tiel and Pankratz, 2021), while studies administering a *sentence* (–world-knowledge) verification task tend to opt for a truth-based paradigm (Bott and Noveck, 2004; De Neys and Schaeken, 2007; Dieussaert et al., 2011; Cremers and Chemla, 2014). It is interesting that this discrepancy rarely seems to be addressed explicitly: Instead, many SPV studies (including those by van Tiel and colleagues—and our own studies here) only mention the actual, felicity-based framing once in the Method section and then, *en passant*, go on to talk about the collected responses as 'True' versus 'False' responses as if they were directly comparable to those from a truth-based verification paradigm. But, of course, this discrepancy does matter: When people behave very differently depending on which of the two possible framings they are presented (as we saw in our pilots), then the first question we should ask ourselves is which of the two framings—if any—elicits responses that are an accurate reflection of the cognitive reality of people's pragmatic processing. It surely cannot be both. (And the tempting question which framing yields more convenient rates of 'pragmatic responses' should ideally only come second.) We have no clear idea how to resolve the first question, let alone a definitive answer to it; but it certainly is an issue that merits to be addressed carefully if the goal is to work with a reliable linking hypothesis between experimental behaviour and the underlying cognitive process of interest.

Regarding the effect pattern on response times predicted by the polarity hypothesis, one should also consider an alternative explanation for cases where it is found: In a seminal paper, Ferreira and Patson (2007) point out that human comprehenders of language will often resort to constructing 'good enough' heuristic representations of linguistic content in their minds as long as they are still so adequate that they can be used to reach the intended goal behind the comprehension process. This is especially the case in artificial situations like reading tasks during psycholinguistic experiments, where the motivation behind comprehension is not really intrinsic to the human subject in the same way that it would be during, say, natural conversation. Crucially, such 'good enough' mental representations of a sentence may fail to grasp all the details of its global constituency or dependency structure. Instead, only salient concepts, particulary those triggered by content words, and, at best, some local morphosyntactic relationships would be stored. It is a reasonable assumption that participants of sentence–picture verification tasks, perhaps especially in the online crowd-sourcing context, may also be prone to sometimes resorting to such 'good enough' parsing strategies while they click themselves through a bunch of dozens of trials so as to finally get to the end and get paid. What unintended consequences could such natural behaviour have, then, in the context of our present main research question? Although a bit speculatively, we would argue that the issue of 'good enough' parsing might introduce an asymmetry between the processing of positively and of negatively polar trials, which, in turn, may be viewed as a (partial) explanation for effect patterns of the kind that the polarity hypothesis predicts. Consider, for example, our critical SPV trials that pertain to the Quantity item:

- *t1+ q*: Some of the **apples** are **red.** & [picture with 16/16 red apples]
- *t1− q*: Not all of the **apples** are **red.** & [picture with 16/16 green apples]

Here, the only two content words featured in each sentence, 'apples' and 'red', have been highlighted by appearing in bold font. This was done to imply that, perhaps, a shallow, 'good enough' parse of either sentence would mainly retain a combination of two lexical concepts APPLE and RED, but not so much the more fine-grained features of each sentence. When evaluating this conceptual combination [APPLE, RED] against a picture showing 16 red apples (trial *t1+ q*), one may feel an instinctive inclination to approve the trial as true/felicitous very quickly. It would take more mental effort, and hence maybe longer time, to reason oneself into full parse of the sentence and to reject it then as false/infelicitous due to pragmatic reasons. By contrast, when evaluating the conceptual combination [Apple, Red] against a picture displaying 16 *green* apples (trial *t1− q*), an intuitive tendency could be to quickly reject the trial as false/infelicitous. Larger cognitive effort, thus more time, would be required to go all the way to a full parse of the underlying sentence in order to be able to reach a literal-interpretation judgement and approve the trial as true/felicitous after all. But all this precisely reflects the average effect pattern on response times that the polarity hypothesis predicts. The only difference is that, given the current assumptions about 'good enough' parsing, it can be explained without recurrence to the alleged nature of mental representations of SIs triggered by diversely polar Horn scales. Of course, one could object to this alternative account by stating that, thanks to exclusion criteria based on response accuracy during control trials, those participants who usually enter statistical analysis will already have been preselected to be very attentive readers who do not resort to 'good enough' parsing strategies at all. But a response to this objection may be that doing well on control trials does not necessarily make someone immune against 'good enough' parsing. In fact, for some types of control trials, 'good enough' parsing even facilitates the correct response (here: *t3*+, *t4*+, *t6*+); and for those types where it does not (here: *t3*−, *t4*−, *t6*−) or where its advantageousness is ambivalent (here: *t2*±, *t5*±), subjects who eventually happen to enter statistical analysis based on accuracy could have just put in a slight additional effort to occasionally go for a full parse indeed. Overall, this small outline of a possible alternative account for the effect pattern of interest calls for future work on this topic to come up with an experimental method that reliably ensures to be sensitive only to the causal relationships implied by the actual account of interest (i.e., the polarity hypothesis and its theoretical justifications).

On the statistical side, there are some issues with the approach of analysing response times (RTs) that we have chosen here. To some extent, this was already discussed in Chapter 6 (particularly, Section 6.3.2): Recall that we first subsetted our data into critical and control trials, then fitted a model on control-trial log RTs, and eventually used that model's predictions as a basis for residualising out variance from the critical-trial log RTs. This is how we ended up with a somewhat artificial dependent measure, *residual log RT*, for critical trials. Now, the problem that we have already discussed in some depth in said chapter (and section) was that due to differences in the control model's random-effects structure across studies, it is a bit difficult to talk about residual log RT as a unitary concept here, since, resultingly, it means something quantitatively slightly different in Study One (Chapter 5) than it does in Study Two (Chapter 6). But aside from this particular problem, there are also broader methodological concerns with the procedure of residualisation as laid out, for example, by York (2012): Among other problematic side effects, residualisation may lead to biased parameter and standard-error estimates. In addition, as Wurm and Fisicaro (2014) stress, it also gives rise to interpretative problems since a residualised variable is essentially a counterfactual abstraction. Thus, perhaps it would indeed have been better for us to simply consider observed log RTs as our dependent measure of interest. The potential nuisance of thereby unintentionally also modelling effects of verification response and polarity that are not unique to critical trials could have been accounted for by including condition (critical/control) and interactions based on it as a further predictor—in a hypothesis-driven model fitted on the entire data set of SPV responses.

Furthermore, in our present experimental studies, we relied on an approach for simplifying the random-effects structure of any of our GLME models in case of non-convergence (incl. singular fit) where we would remove random-slope parameters in an arbitrary prespecified order until convergence was reached. However, this is not an ideal approach in the sense that it might yield a random-effects structure that is actually not the most complex one for which convergence is still ensured. For example, following our approach, we would always remove by-item main-effect slopes first and only then by-subject main-effect slopes. Yet, by doing this, we could easily 'miss out on' a perfectly converging model that still includes by-item main-effect slopes, but no by-subject main-effect slopes. Likewise, our approach did not foresee to probe whether constraining certain random intercept–slope or slope–slope correlation terms could potentially make some model converge properly. In summary, a more informed preregistered line of action could have been to systematically remove random slopes and random correlation terms in such an order that the parameters which account for the least amount of variance are dropped first, hence roughly following recommendations given, e.g., by Bates, Kliegl, Vasishth and Baayen (2015).

Lastly, another aspect of our preregistered, hypothesis-driven analyses which may have been suboptimal was not to consider order of trial presentation as a covariate predictor there. As we clearly observed in some of our exploratory analyses, in both Chapters 5 and 6, order of trial presentation consistently appeared to be the most salient predictor of SPV response times. So, the fact that our hypothesis-driven models were fitted without it may have rendered the estimation of their remaining parameters, including predictors of interest, more noisy than it had needed to be.

Chapter 8
Conclusion

The goal of the present work was to acquire a better understanding of factors that modulate pragmatic processing in human comprehenders. In particular, we have been concerned with scalar implicatures (SIs). Within that context, we examined if a recently emerged view on SIs (the polarity hypothesis) and its predictions about human response behaviour in SI-eliciting verification tasks hold up to experimental scrutiny. In addition, we assessed individual differences in working memory capacity, fluid intelligence, and print exposure in an exploratory attempt to reveal possible relationships between such differences and aspects of SI processing behaviour.

Based on evidence from two experimental studies ($N = 100$; $N = 400$) that we conducted through web-based crowd-sourcing, we have eventually obtained only mild support for the polarity hypothesis: The predicted effect pattern on response times in a sentence–picture verification (SPV) task was manifest significantly in our second, but not our first, smaller-sample study. However, even there where it was, in fact, present, closer examination revealed it to exclusively depend on the results for one experimental item in particular, featuring the scalar term 'might (not) be', but to be unsupported by all three other tested items. This calls for future research to re-evaluate the polarity hypothesis more closely, especially in regards to potential between-scale differences that could be modulating the behavioural effects that it predicts.

As for individual differences, which were only analysed in our first study ($N = 100$), our exploratory findings suggest that higher fluid intelligence (measured by a Raven's Progressive Matrices task) is associated with an increase in response duration on pragmatically ambiguous SPV trials as compared to unambiguous ones. Possibly, this indicates that more intelligent people are rather sensitive to the response dilemma caused by pragmatic ambiguity, thus leading them to display a response slowdown. Besides, we also observe a rather intricate three-way interaction on response times between print exposure (measured by an author recognition test), verification response, and polarity. But due to its complexity, its lower degree of significance, and the multitude of comparisons that were performed, it remains unclear if that three-way interaction does, indeed, reflect a meaningful and cognitively real relationship or not. In any case, future work is invited to try to confirm both of these potential effects of individual differences within a confirmatory study design. Lastly, despite promising hints from previous literature, we did not find working memory capacity (measured by an OSpan task) to bear any responsibility in accounting for variation in either response times or response choices given by participants in the SPV task.

To conclude, while it **might be** the case that **sometimes** people's processing of scalar implicatures is affected by polarity, there is also reason to believe that **not all** kinds of people are equally sensitive to pragmatic ambiguity. Keep in mind, though, that surely **not everywhere** in the real world one will see language comprehension taking place exactly as it does in an experimental set-up.

Accessible Data and Code

All utilised experiment files, collected response data, and R scripts for statistical analysis can be found in these two OSF repositories, concerning our two main studies:

- **osf.io/c52vp** (Study One: Project repository)
- **osf.io/w49sz** (Study Two: Project repository)

Furthermore, each of the two main studies was officially preregistered on OSF. The corresponding preregistration documentation pages are the following:

- **osf.io/pzja3** (Study One: Preregistration)
- **osf.io/dhpzq** (Study Two: Preregistration)

A post-data-collection addendum statement to the preregistration protocol of Study Two has been made available under the following URL:

- **osf.io/zrveq** (Study Two: Addendum to preregistration)

Regarding our three conducted pilot studies, note that their experiment files, response data, and analysis scripts are stored separately in the following OSF repositories:

- **osf.io/qr6bs** (Pilot 28 Jan. 2023: Project repository)
- **osf.io/ayj54** (Pilot 13 Feb. 2023: Project repository)
- **osf.io/mqpk2** (Pilot 16 Feb. 2023: Project repository)

To test different versions of the experimental tasks on a local machine, the Lingoturk software (Pusse et al., 2016) needs to be installed. It can be downloaded from here:

- **github.com/FlorianPusse/Lingoturk**

Alternatively, final versions of the experimental tasks remain easy to access, probably at least for a couple of weeks after the submission of the present thesis, via the following URLs where they are hosted on servers belonging to Saarland University:

1. Sentence–Picture Verification (SPV):
    - For isolated administration:
        - **multitude.coli.uni-saarland.de:8080/renderProlific?expId=4681**
        - **masses.coli.uni-saarland.de:8080/renderProlific?expId=3961**
    - For administration conjointly with ID tasks:
        - **multitude.coli.uni-saarland.de:8080/renderProlific?expId=4661**
        - **masses.coli.uni-saarland.de:8080/renderProlific?expId=3946**
2. Operation Span (OSpan):
    - **multitude.coli.uni-saarland.de:8080/renderProlific?expId=4662**
    - **masses.coli.uni-saarland.de:8080/renderProlific?expId=3948**

3. Author Recognition Test (ART):
   - **multitude.coli.uni-saarland.de:8080/renderProlific?expId=4663**
   - **masses.coli.uni-saarland.de:8080/renderProlific?expId=3945**
4. Raven's Progressive Matrices (RPM):
   - **multitude.coli.uni-saarland.de:8080/renderProlific?expId=4664**
   - **masses.coli.uni-saarland.de:8080/renderProlific?expId=3947**

Last but not least, a summary of links to (differently randomised) versions of the pre-testing questionnaire that was filled out by 20 native English speakers back in December 2022 and featured prototypes of the SPV (control) trials can be found here:

- **osf.io/796nx**

The questionnaire responses that were eventually collected have been stored in the following data file:

- **osf.io/kt9z5**

Some more information on that pre-test is provided right below in Appendix A.

Appendix A
Pre-Test Design and Results

In December 2022, a pre-test of the linguistic and visual materials for the sentence–picture verification (SPV) task was carried out. It featured prototypical versions of said materials, and its goal was to find out if native English speakers would, indeed, comprehend the materials as we had intended.

The set-up basically looked like this: Each participant of the pre-test would receive a link to a unique version of our pre-testing questionnaire (implemented in Google Forms). In the questionnaire, they would first have to provide basic demographic information about themselves, particularly an assertion about indeed being a native English speaker. Then, the main part of the questionnaire would begin, where 40 different combinations of a sentence and a picture would be displayed. For each participant, these 40 combinations were ordered in a uniquely randomised way—hence why the unique versions of the questionnaire. Right below each such sentence–picture combination, participants were asked to select which of three options would best describe the combination according to them: (1) 'The statement is TRUE.' (2) 'The statement is FALSE.' (3) 'I am not sure.'—Further, a text field was provided additionally where participants were encouraged to describe the reasoning behind their selection in their own words, especially if they had opted for the third, indecisive option. Crucially, the questionnaire featured only combinations that were structurally equivalent to what we call *control trials* later on in our two main experimental studies (i.e., of trial types *t2*...*t6*± in Table 4.3). That is, all these combinations were designed to, ideally, each be interpretable only in one unambiguous way (as either true or false), with no SI-based pragmatic ambiguity being built in anywhere. This was because we wanted this pre-test to be a sanity check if native speakers are, at least, in agreement with us and among each other as to what the correct responses to control trials are—only once that is ensured, it would make sense at all to also measure responses to potentially SI-inducing sentence–picture combinations derived from the same materials.

In Table A.1, the sentences that were presented to participants of the pre-test have been listed. As can be noticed, they resemble—but are not identical to—the sentences eventually used in the pilot and main experimental studies (cf. Table 4.2). They are, however, structured in the same way as our final sentence items, with four prototypical

Table A.1: Pre-testing prototypes of the four sentence items for the sentence–picture verification task in all their possible variants. Compare these to their final versions which are listed in Table 4.2 and which were eventually employed in our main experimental studies.

Item	Scalar Type	Sentence
Quantity	∃	Some of the apples are green.
(Prototype)	∀	All of the apples are green.
	¬ ∀	Not all of the apples are green.
	¬ ∃	None of the apples are green.
Possibility	∃	The arrow might land on a blue segment.
(Prototype)	∀	The arrow will definitely land on a blue segment.
	¬ ∀	The arrow might not land on a blue segment.
	¬ ∃	The arrow will definitely not land on a blue segment.
Time	∃	She sometimes hit the bullseye.
(Prototype)	∀	She always hit the bullseye.
	¬ ∀	She did not always hit the bullseye.
	¬ ∃	She never hit the bullseye.
Space	∃	It is raining somewhere in the country.
(Prototype)	∀	It is raining everywhere in the country.
	¬ ∀	It is not raining everywhere in the country.
	¬ ∃	It is not raining anywhere in the country.

items labelled Quantity, Possibility, Time, and Space where each item encompasses four different sentence variants based on scalar types ∃, ∀, ¬ ∀, and ¬ ∃.

As for the corresponding pictures, they were identical to their final versions in case of the prototypical Time item (see back in Figure 4.3), but there were differences to the final versions in case of the remaining three items: While the prototypical Quantity item featured the exact same drawings as its final version, shown in Figure 4.1, the 0 % and 100 % picture variants had still been reversed, i.e., in the pre-test, it was the 100 % variant that showed 16/16 *green* apples, while the 0 % variant showed 16/16 *red* apples. Analogously, the question under discussion in the prototypical Quantity item's sentences was all about whether apples are 'green' (rather than 'red', as in the final version); we will see in a moment how this eventual reversal of colours was motivated. Moving on, instead of showing a Galton board with a falling bead as in its final version (Figure 4.2), the prototype of the Possibility item rather displayed a kind of fortune wheel as shown in Figure A.1, where the wheel was drawn as such that it appeared to be in motion and where it was divided into eight segments that could be either of blue or of yellow colour. Importantly, a pointing arrow was placed atop of the fortune wheel, and the picture variants differed in whether all segments were of uniform colour (i.e., either all blue or all yellow) or if there rather were four blue and four yellow, intermixed segments. Lastly, the prototype of the Space item featured pictures showing a simple map of the United Kingdom, on top of which four symbolic icons representing either sunny weather or rain had been drawn. The picture variants, presented in Figure A.2, differed in how many of the icons represented rain (i.e., either none, half, or all of them).

Please find links to all administered versions of the pre-testing questionnaire in the following document: **osf.io/796nx**. Overall, 20 native speakers of English were successfully recruited for participation in the pre-test by snowball-sampling through personal ac-

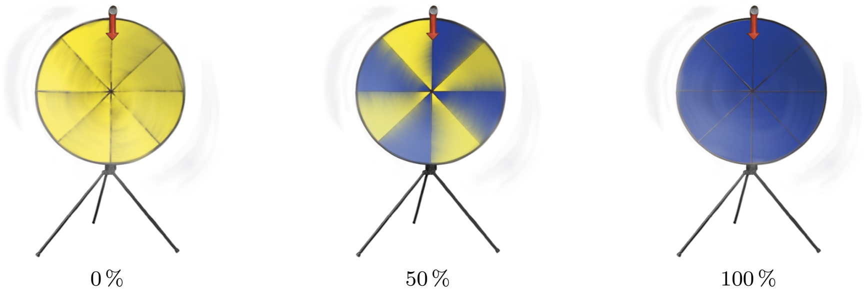


Figure A.1: Pre-testing prototypes of the three picture variants for the Possibility item. The percentages annotated below indicate the gradual degree of informative fulfilment based on the positive scale. Compare these to their final versions which are shown in Figure 4.2 and which were eventually employed in our main experimental studies.

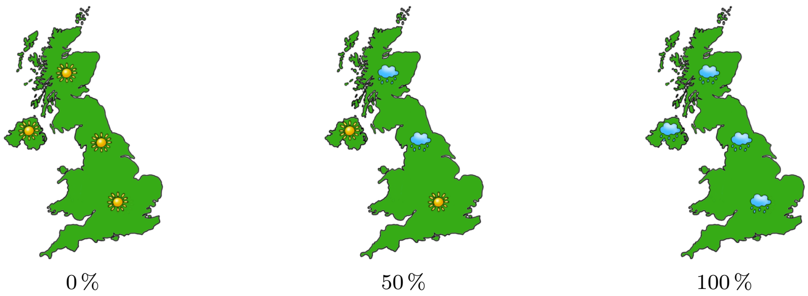


Figure A.2: Pre-testing prototypes of the three picture variants for the Space item. The percentages annotated below indicate the gradual degree of informative fulfilment based on the positive scale. Compare these to their final versions which are shown in Figure 4.4 and which were eventually employed in our main experimental studies.

quaintances. Their average age was 32 years, with a standard deviation of 15, while the youngest participant was 15 years old and the oldest one 70 years. Self-reported genders were 14 times ‘male’, 5 times ‘female’, and once ‘non-binary’. Countries of origin ranged from the United States (10 participants), Canada (1 participant), Ireland (1 participant), and the United Kingdom (6 participants) all the way to Bangladesh (1 participant) and New Zealand (1 participant).

Results

The collected pre-testing response data can be accessed here: **osf.io/kt9z5**. All responses were received between 9 and 14 December 2022. Out of the 20 participants, there was one male subject (subject ID `"S15"` in the data file) who particularly stood out because his number of trials responded to as expected was only 17 (out of 40), thereby more than 3 standard deviations below the average performance of the remaining participants. Crucially, however, the qualitative feedback provided by that subject is very relevant as it sheds light on the main reason for his poor performance. As he indicates, he is colour-blind, which explains his inability to answer reliably on the Quantity (red vs. green apples) and Possibility (yellow vs. blue segments) items. This insight was actually what motivated us to require normal vision without colour-blindness among the recruiting criteria for our proper experimental studies on Prolific afterwards (see Section 4.2).

Looking just at the 19 remaining, non-colour-blind participants, their mean accuracies in terms of alignment with our expected responses were the following for each item: 99.5 % for Quantity, 92.6 % for Possibility, 96.8 % for Time, and 86.8 % for Space. What was more informative for us than these quantitative summaries, however, was a lot of the qualitative feedback in form of text, provided by several participants—such qualitative feedback was eventually what motivated our final modifications to the sentence–picture items ahead of conducting our proper experiments on Prolific. We give a brief summary of that, here:

One participant highlighted that the word 'green' is ambiguous in relation to apples as it can indicate either their colour or their ripeness. Thus, some native English speakers may find it plausible that a red(-colour) apple is also green (= unripe). The issue of potential distortions caused by this lexical ambiguity was thus fixed by switching the colour of reference from 'green' to 'red' in the final version of the Quantity item: That is, in our actual Prolific experiments, the employed sentences were these: '(Some|All|Not|Not all) of the apples are **red.**' Accordingly, the 100 % picture variant that we finally opted for in the actual experiments was one that showed only red apples.

Regarding the Time item, several participants noted that sentences like 'She (always|never) the bullseye.' feel to them as if they referred to the displayed archer's whole lifetime rather than just her success in the short-term situation visible on the picture. Therefore, for the actual experiments, we added the adverbial specification '[...] today.' to the sentences of the Time item in an attempt to disambiguate. Hence we finally used sentences of this type: 'She (sometimes|always|did not always|never) hit the bullseye **today.**'

With respect to the Possibility item's prototype, many participants expressed confusion especially about picture variants showing a fortune wheel with only yellow (or only blue) segments. In other words, it seemed odd to them, e.g., to have the possibility of an arrow landing on a 'blue segment' under discussion, while there was no visible blue segment whatsoever on the picture to relate the sentence to. Some also commented speculating if the colour of the thin lines between visual segments played any role. Other subjects had a problem with giving a verification response to future-tense sentences, even despite the adverbial specification 'definitely' in the expression 'will definitely (not)'. Overall, this item appeared to be quite problematic in its prototypical form. That is why, eventually, we replaced it entirely by a drastically altered final version, featuring completely new sentences and pictures, i.e., the ones with the Galton board and the falling bead. Thereby, the issues brought up by the pre-test participants were likely resolved.

The Space item's prototype was the most problematic one for many participants: Here, it seemed odd for many of them to verify a generalised statement like 'It is raining everywhere in the country.' while only seeing sunny-weather or rain icons in four very specific locations on the map of the UK. Secondly, several subjects noted that they perceived the variant 'It is not raining everywhere in the country.' as ambiguous in terms of the scope of the negation: It may mean either NOT[EVERYWHERE[RAINING]] (= intended meaning) or EVERYWHERE[NOT[RAINING]] (= unintended meaning). Moreover, one participant noted that the expression 'the country' may be ambiguous too, considering that it can refer either to the entirety of the UK or only to one of its broader regions (i.e., countries) England, Wales, Scotland, or Northern Ireland. With all of this considered, we came up with a linguistic and visual solution to these discussed issues by eventually substituting said prototype with a new and final version of the Space item, now being about Africa (rather than the UK) and daytime (rather than rain) as its space and property of reference, respectively, and where the adverbial scalar had been moved into a topicalised position at the very beginning of a sentence.

Appendix B
Study One: Additional Figures/Tables

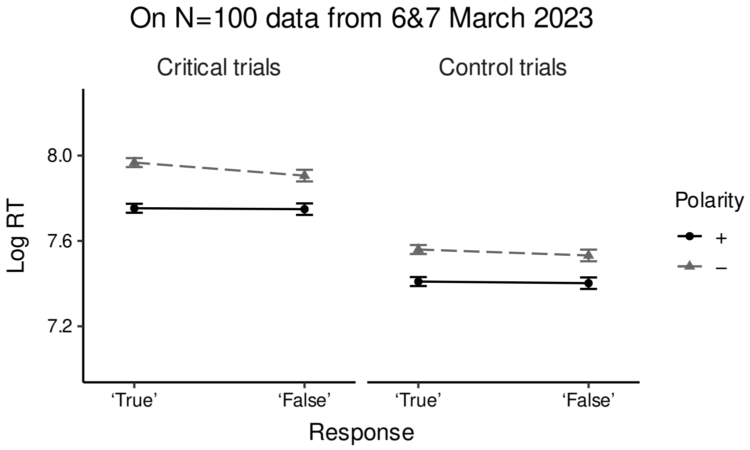


Figure B.1: Mean log RTs grouped by condition, by verification response, and by polarity. The displayed 95 % CIs are based on within-subject-and-item standard errors (computed following Morey et al., 2008). The underlying SPV response data was collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). One is invited to compare this plot to a sister plot in Section 5.3.6 (Figure 5.2) where residual log RTs rather than log RTs are displayed based on the same data. Moreover, there is a twin plot further below in Appendix C where analogous results from a follow-up study with larger sample size are presented: Figure C.1.

Table B.1: Random-effect estimates from linear mixed-effects Model (i), fitted to residual log RT, based on SPV response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). Here, $\hat{\tau}$ refers to parameter estimates of group-level standard deviations (by-subject or by-item), while any $\hat{\rho}_{\ldots,\ldots}$ refers to parameter estimates of intercept–slope or slope–slope correlation terms. Wherever the mention of an estimate's value would either be redundant (cases $\hat{\rho}_{\mathrm{A,B}} = \hat{\rho}_{\mathrm{B,A}}$ and $\hat{\rho}_{\mathrm{A,A}} = 1$) or impossible because the parameter in question was constrained to zero in order to achieve model convergence, a placeholder '—' is shown instead. Corresponding fixed-effect estimates from the same model are summarised above in Section 5.3.6, in Table 5.1. Moreover, this table has a twin further below in Appendix C where analogous results from a follow-up study with larger sample size are reported: Table C.1.

Groups	Term	$\hat{\tau}$	$\hat{\rho}_{\mathtt{1},\ldots}$	$\hat{\rho}_{\mathtt{VR},\ldots}$	$\hat{\rho}_{\mathtt{POL},\ldots}$
Subject	Intercept (`1`)	0.12	—	—	—
	Verification response (`VR`)	0.11	−0.06	—	—
	Polarity (`POL`)	0.09	−0.38	0.55	—
	Interaction of `VR:POL`	—	—	—	—
Item	Intercept (`1`)	0.09	—	—	—
	Verification response (`VR`)	—	—	—	—
	Polarity (`POL`)	0.07	−0.84	—	—
	Interaction of `VR:POL`	—	—	—	—

Table B.2: Random-effect estimates from linear mixed-effects Model (ii), fitted to (critical-trial) verification response, based on SPV response data collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). Here, $\hat{\tau}$ refers to parameter estimates of group-level standard deviations (by-subject or by-item), while $\hat{\rho}_{\mathtt{1},\mathtt{POL}}$ refers to parameter estimates of intercept–slope correlation terms. Corresponding fixed-effect estimates from the same model are summarised above in Section 5.3.6, in Table 5.2.

Groups	Term	$\hat{\tau}$	$\hat{\rho}_{\mathtt{1},\mathtt{POL}}$
Subject	Intercept (`1`)	1.45	—
	Polarity (`POL`)	0.87	0.31
Item	Intercept (`1`)	0.78	—
	Polarity (`POL`)	1.68	0.76

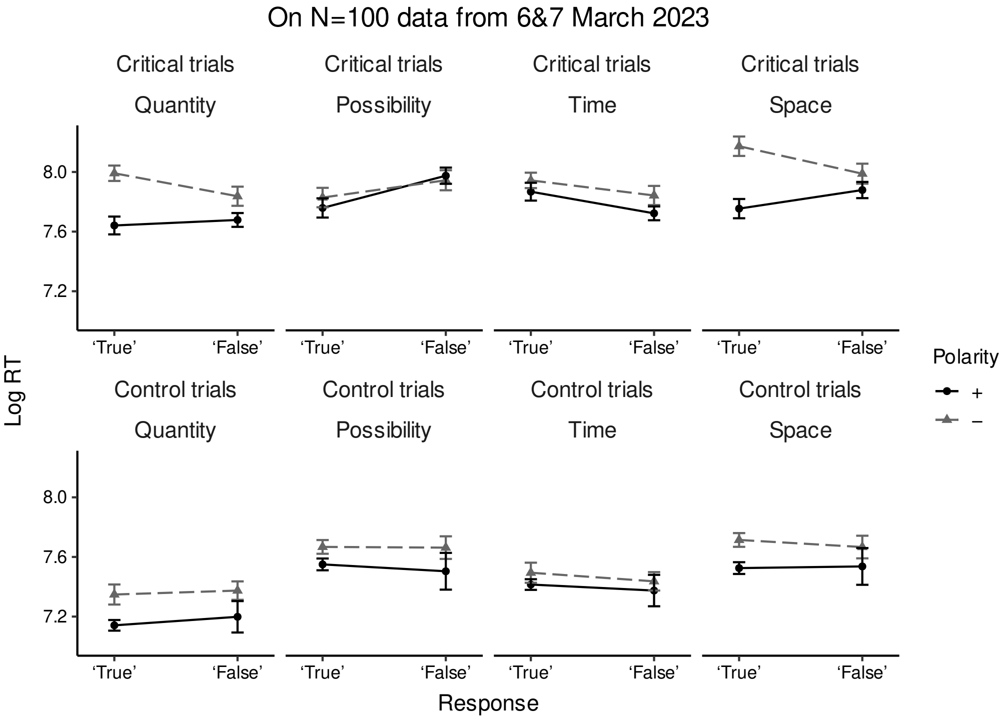


Figure B.2: For each item separately: Mean log RTs grouped by condition, by verification response, and by polarity. The displayed error bars represent within-subject standard errors (computed following Morey et al., 2008). The underlying SPV response data was collected on 6 and 7 March 2023 from 100 Prolific participants (post-exclusion). One is invited to compare this plot to a sister plot in Section 5.3.6 (Figure 5.3) where residual log RTs rather than log RTs are displayed based on the same data. Moreover, there is a twin plot further below in Appendix C where analogous results from a follow-up study with larger sample size are presented: Figure C.2.

Appendix C
Study Two: Additional Figures/Tables

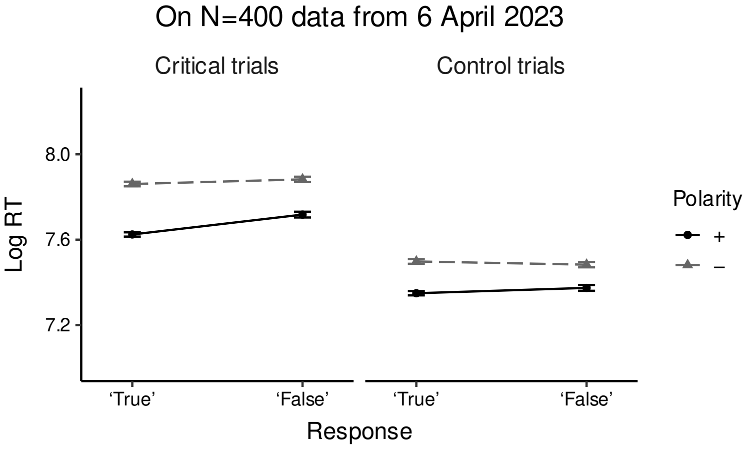


Figure C.1: Mean log RTs grouped by condition, by verification response, and by polarity. The displayed error bars represent within-subject-and-item standard errors (computed following Morey et al., 2008). The underlying SPV response data was collected on 6 April 2023 from 400 Prolific participants (post-exclusion). One is invited to compare this plot to a sister plot in Section 6.3.2 (Figure 6.3) where residual log RTs rather than log RTs are displayed based on the same data. Moreover, there is a twin plot above in Appendix B where analogous results from our previous study with smaller sample size are presented: Figure B.1.

Table C.1: Random-effect estimates from linear mixed-effects New Model (i), fitted to residual log RT, based on SPV response data collected on 6 April 2023 from 400 Prolific participants (post-exclusion). Here, $\hat{\tau}$ refers to parameter estimates of group-level standard deviations (by-subject or by-item), while any $\hat{\rho}_{\ldots,\ldots}$ refers to parameter estimates of intercept–slope or slope–slope correlation terms. Wherever the mention of an estimate's value would either be redundant (cases $\hat{\rho}_{\mathrm{A,B}} = \hat{\rho}_{\mathrm{B,A}}$ and $\hat{\rho}_{\mathrm{A,A}} = 1$) or impossible because the parameter in question was constrained to zero in order to achieve model convergence, a placeholder '—' is shown instead. Corresponding fixed-effect estimates from the same model are summarised above in Section 6.3.2, in Table 6.2. Moreover, this table has a twin further below in Appendix B where analogous results from our previous study with smaller sample size are reported: Table B.1. Note that this table looks surprisingly empty because our eventually converging variant of New Model (i) featured only by-subject and by-item random intercepts, but no random slopes whatsoever.

Groups	Term	$\hat{\tau}$	$\hat{\rho}_{\mathtt{1},\ldots}$	$\hat{\rho}_{\mathtt{VR},\ldots}$	$\hat{\rho}_{\mathtt{POL},\ldots}$
Subject	Intercept (`1`)	0.14	—	—	—
	Verification response (`VR`)	—	—	—	—
	Polarity (`POL`)	—	—	—	—
	Interaction of `VR:POL`	—	—	—	—
Item	Intercept (`1`)	0.14	—	—	—
	Verification response (`VR`)	—	—	—	—
	Polarity (`POL`)	—	—	—	—
	Interaction of `VR:POL`	—	—	—	—

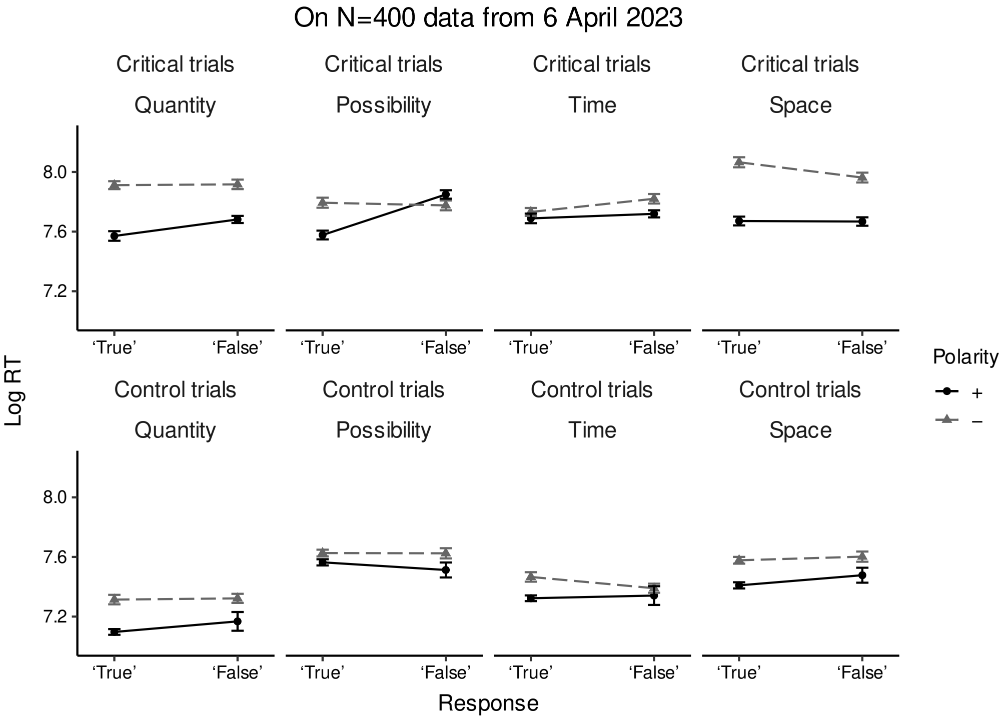


Figure C.2: For each item separately: Mean log RTs grouped by condition, by verification response, and by polarity. The displayed error bars represent within-subject standard errors (computed following Morey et al., 2008). The underlying SPV response data was collected on 6 April 2023 from 400 Prolific participants (post-exclusion). One is invited to compare this plot to a sister plot in Section 6.3.2 (Figure 6.4) where residual log RTs rather than log RTs are displayed based on the same data. Moreover, there is a twin plot above in Appendix B where analogous results from our previous study with smaller sample size are presented: Figure B.2.

Bibliography

Acheson, D. J., Wells, J. B. and MacDonald, M. C. (2008), ‘New and updated tests of print exposure and reading abilities in college students’, *Behavior research methods* **40**(1), 278–289.

Atlas, J. D. and Levinson, S. C. (1981), It-clefts, informativeness and logical form: Radical pragmatics (revised standard version), *in* ‘Radical pragmatics’, Academic Press, pp. 1–62.

Attridge, N. and Inglis, M. (2014), ‘Intelligence and negation biases on the conditional inference task: A dual-processes analysis’, *Thinking & reasoning* **20**(4), 454–471.

Bates, D., Kliegl, R., Vasishth, S. and Baayen, H. (2015), ‘Parsimonious mixed models’, arXiv preprint arXiv:1506.04967.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015), ‘Fitting linear mixed-effects models using lme4’, *Journal of statistical software* **67**(1), 1–48.

Bill, C., Romoli, J. and Schwarz, F. (2018), ‘Processing presuppositions and implicatures: Similarities and differences’, *Frontiers in communication* **3**, 44.

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

Buccola, B., Križ, M. and Chemla, E. (2022), ‘Conceptual alternatives: Competition in language and beyond’, *Linguistics and philosophy* **45**(2), 265–291.

Bürkner, P.-C. (2017), ‘brms: An R package for Bayesian multilevel models using Stan’, *Journal of statistical software* **80**, 1–28.

Cattell, R. B. (1963), ‘Theory of fluid and crystallized intelligence: A critical experiment’, *Journal of educational psychology* **54**(1), 1.

Chierchia, G. et al. (2004), ‘Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface’, *Structures and beyond* **3**, 39–103.

Clark, H. H. and Chase, W. G. (1972), ‘On the process of comparing sentences against pictures’, *Cognitive psychology* **3**(3), 472–517.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O. and Engle, R. W. (2005), ‘Working memory span tasks: A methodological review and user’s guide’, *Psychonomic bulletin & review* **12**(5), 769–786.

Cremers, A. and Chemla, E. (2014), Direct and indirect scalar implicatures share the same processing signature, *in* ‘Pragmatics, semantics and the case of SIs’, Springer, pp. 201–227.

De Neys, W. and Schaeken, W. (2007), ‘When people are more logical under cognitive load: Dual task impact on scalar implicature’, *Experimental psychology* **54**(2), 128.

Degen, J. (2015), 'Investigating the distribution of some (but not all) implicatures using corpora and web-based methods', *Semantics and pragmatics* **8**, 11.

Degen, J. and Tanenhaus, M. K. (2015), 'Processing scalar implicature: A constraint-based approach', *Cognitive science* **39**(4), 667–710.

Deutsch, R., Kordts-Freudinger, R., Gawronski, B. and Strack, F. (2009), 'Fast and fragile: A new look at the automaticity of negation processing.', *Experimental psychology* **56**(6), 434–446.

Dieussaert, K., Verkerk, S., Gillard, E. and Schaeken, W. (2011), 'Some effort for some: Further evidence that scalar implicatures are effortful', *Quarterly journal of experimental psychology* **64**(12), 2352–2367.

Doran, R., Baker, R., McNabb, Y., Larson, M. and Ward, G. (2009), 'On the non-unified nature of scalar implicature: An empirical investigation', *International review of pragmatics* **1**(2), 211–248.

Feeney, A., Scrafton, S., Duckworth, A. and Handley, S. J. (2004), 'The story of some: Everyday pragmatic inference by children and adults', *Canadian journal of experimental psychology/Revue canadienne de psychologie expérimentale* **58**(2), 121–132.

Ferreira, F. and Patson, N. D. (2007), 'The "good enough" approach to language comprehension', *Language and linguistics compass* **1**(1-2), 71–83.

Fodor, J. A. and Garrett, M. F. (1975), 'The psychological unreality of semantic representations', *Linguistic inquiry* **6**(4), 515–531.

Fox, D. and Katzir, R. (2011), 'On the characterization of alternatives', *Natural language semantics* **19**, 87–107.

Gazdar, G. (1979), 'Pragmatics, implicature, presuposition and logical form', *Critica* **12**(35), 113–122.

Gelman, A. and Carlin, J. (2014), 'Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors', *Perspectives on psychological science* **9**(6), 641–651.

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

Greaney, V. (1980), 'Factors related to amount and type of leisure time reading', *Reading research quarterly* **15**(3), 337–357.

Green, P. and MacLeod, C. 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.

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

Hirschberg, J. L. B. (1991), *A theory of scalar implicature*, Garland Publishing.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian journal of statistics* **6**(2), 65–70.

Horn, L. R. (1972), *On the semantic properties of logical operators in English*, University of California, Los Angeles.

Hu, J., Levy, R. and Schuster, S. (2022), Predicting scalar diversity with context-driven uncertainty over alternatives, *in* 'Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics', pp. 68–74.

Huang, Y. T., Spelke, E. and Snedeker, J. (2013), 'What exactly do numbers mean?', *Language learning and development* **9**(2), 105–129.

Johnson, E. and Arnold, J. E. (2021), 'Individual differences in print exposure predict use of implicit causality in pronoun comprehension and referential prediction', *Frontiers in psychology* **12**, 2933.

Just, M. A. and Carpenter, P. A. (1992), 'A capacity theory of comprehension: individual differences in working memory.', *Psychological review* **99**(1), 122.

Kissine, M. and De Brabanter, P. (2023), 'Pragmatic responses to under-informative some-statements are not scalar implicatures', *Cognition* **237**, 105463.

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

Lewandowski, D., Kurowicka, D. and Joe, H. (2009), 'Generating random correlation matrices based on vines and extended onion method', *Journal of multivariate analysis* **100**(9), 1989–2001.

Li, E., Schuster, S. and Degen, J. (2021), Predicting scalar inferences from "or" to "not both" using neural sentence encoders, *in* 'Proceedings of the Society for Computation in Linguistics 2021', pp. 446–450.

Martin-Chang, S. L. and Gould, O. N. (2008), 'Revisiting print exposure: Exploring differential links to vocabulary, comprehension and reading rate', *Journal of research in reading* **31**(3), 273–284.

Marty, P., Chemla, E. and Spector, B. (2013), 'Interpreting numerals and scalar items under memory load', *Lingua* **133**, 152–163.

Marty, P., Romoli, J., Sudo, Y., van Tiel, B. and Breheny, R. (2020), 'Processing implicatures: A comparison between direct and indirect SIs', Oral presentation at Experiments in Linguistic Meaning (ELM), Philadelphia, PA.

Mayn, A. and Demberg, V. (2022), Individual differences in a pragmatic reference game, *in* 'Proceedings of the Annual Meeting of the Cognitive Science Society', Vol. 44, pp. 3016–3022.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A. and Wager, T. D. (2000), 'The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis', *Cognitive psychology* **41**(1), 49–100.

Morey, R. D. et al. (2008), 'Confidence intervals from normalized data: A correction to Cousineau (2005)', *Tutorials in quantitative methods for psychology* **4**(2), 61–64.

Pusse, F., Sayeed, A. and Demberg, V. (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', pp. 57–61.

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Raven, J. C. (1938), 'Standardization of progressive matrices', *British journal of medical psychology* **19**(1), 137–150.

Romoli, J. and Schwarz, F. (2015), An experimental comparison between presuppositions and indirect scalar implicatures, *in* 'Experimental perspectives on presuppositions', Springer, pp. 215–240.

Ronai, E. and Xiang, M. (2022), 'Quantifying semantic and pragmatic effects on scalar diversity', *Proceedings of the Linguistic Society of America* **7**(1), 5216.

Ryzhova, M., Mayn, A. and Demberg, V. (2023), What inferences do people actually make upon encountering informationally redundant utterances? An individual differences study, *in* 'Proceedings of the Annual Meeting of the Cognitive Science Society', Vol. 45.

Satterthwaite, F. E. (1946), 'An approximate distribution of estimates of variance components', *Biometrics bulletin* **2**(6), 110–114.

Scholman, M. C., Demberg, V. and Sanders, T. J. (2020), 'Individual differences in expecting coherence relations: Exploring the variability in sensitivity to contextual signals in discourse', *Discourse processes* **57**(10), 844–861.

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

Schweizer, K. (2007), 'Investigating the relationship of working memory tasks and fluid intelligence tests by means of the fixed-links model in considering the impurity problem', *Intelligence* **35**(6), 591–604.

Scribner, S. and Cole, M. (1981), 'Unpackaging literacy', *Writing: The nature, development, and teaching of written communication* **1**, 71–87.

Spector, B. (2013), 'Bare numerals and scalar implicatures', *Language and linguistics compass* **7**(5), 273–294.

Sperber, D. and Wilson, D. (1986), *Relevance: Communication and cognition*, Blackwell.

Staab, J., Urbach, T. P. and Kutas, M. (2008), 'Negation processing in context is not (always) delayed', *Center for research in language* **20**(3), 3–34.

Stanovich, K. E. and West, R. F. (1989), 'Exposure to print and orthographic processing', *Reading research quarterly* **24**(4), 402–433.

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

Turner, M. L. and Engle, R. W. (1989), 'Is working memory capacity task dependent?', *Journal of memory and language* **28**(2), 127–154.

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

van Tiel, B., Pankratz, E. and Sun, C. (2019), 'Scales and scalarity: Processing scalar inferences', *Journal of memory and language* **105**, 93–107.

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

Vasishth, S. (2023), 'Some right ways to analyze (psycho)linguistic data', *Annual review of linguistics* **9**, 273–291.

Wald, A. (1943), 'Tests of statistical hypotheses concerning several parameters when the number of observations is large', *Transactions of the American Mathematical Society* **54**(3), 426–482.

Wang, S., Sun, C., Tian, Y. and Breheny, R. (2021), 'Verifying negative sentences', *Journal of psycholinguistic research* **50**(6), 1511–1534.

Whelan, R. (2008), 'Effective analysis of reaction time data', *The psychological record* **58**(3), 475–482.

Wurm, L. H. and Fisicaro, S. A. (2014), 'What residualizing predictors in regression analyses does (and what it does not do)', *Journal of memory and language* **72**, 37–48.

Yang, X., Minai, U. and Fiorentino, R. (2018), 'Context-sensitivity and individual differences in the derivation of scalar implicature', *Frontiers in psychology* **9**, 1720.

York, R. (2012), 'Residualization is not the answer: Rethinking how to address multicollinearity', *Social science research* **41**(6), 1379–1386.