The influence of polarity items on inferential judgments

Denić, M.; Homer, V.; Rothschild, D.; Chemla, E.

DOI
10.1016/j.cognition.2021.104791
Publication date
2021
Document Version
Final published version
Published in
Cognition

Cognition 215 (2021) 104791

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

The influence of polarity items on inferential judgments

Milica Denić [a,*], Vincent Homer [c,e], Daniel Rothschild [d], Emmanuel Chemla [b]

[a] Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
[b] Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Études Cognitives, École Normale Supérieure, PSL University, France
[c] Department of Linguistics, University of Massachusetts at Amherst, United States
[d] Department of Philosophy, University College London, United Kingdom
[e] Institut Jean-Nicod (ENS, EHESS, CNRS), Département d'Études Cognitives, École Normale Supérieure, PSL University, France

ARTICLE INFO

Keywords:
Modularity
Polarity
Monotonicity
Intuitions
Reasoning

ABSTRACT

Polarity items are linguistic expressions such as *any*, *at all*, *some*, which are acceptable in some linguistic environments but not others. Crucially, whether a polarity item is acceptable in a given environment is argued to depend on the inferences (in the reasoning sense) that this environment allows. We show that the inferential judgments reported for a given environment are modified in the presence of polarity items. Hence, there is a two-way influence between linguistic and reasoning abilities: the linguistic acceptability of polarity items is dependent on reasoning facts and, conversely, reasoning judgments can be altered by the mere addition of seemingly innocuous polarity items.

1. Polarity items and monotonicity

Monotonicity is an abstract, logical property that a linguistic environment is said to have when this environment systematically supports inferences from subsets to supersets or vice versa. For instance, an environment is *upward monotone* or *upward-entailing* (UE) if it supports a subset to superset inference; an example is the environment of the boldface expressions in (1). Similarly, a *downward monotone* or *downward-entailing* (DE) environment supports the superset to subset inference; an example is in (2).

(1) This animal is **a siamese cat**.

⇝ This animal is **a cat**.

(2) This animal isn't **a cat**.

⇝ This animal isn't **a siamese cat**.

Interestingly, there is a class of expressions, called *polarity items* (PIs) whose acceptability has been linked to the logical property of monotonicity. This was first proposed by Fauconnier (1975) and Ladusaw (1979), in relation to the most studied category of such expressions, namely negative polarity items (NPIs) such as *any*, *ever*, and *at all*. The generalization proposed for the distribution of NPIs is that they are acceptable *(=licensed)* in a DE environment, as in (3), and not acceptable in an UE environment, as in (4).

(3) This animal isn't a cat **at all**.
(4) animal is a cat **at all**.

In addition, there are *positive polarity items* (PPIs) such as *some*, *something*, *someone* that are acceptable in UE environments as in (5), but that cannot be interpreted in a number of DE environments. For instance, (6) doesn't have a reading in which the existential quantifier *some coffee* is interpreted with narrow scope with respect to negation (i.e. a reading equivalent to 'I didn't drink any coffee'). It is the lack of narrow scope of *some* under negation which signals that it is a PPI. Note that the sentence is acceptable under a reading where *some* outscopes negation (i.e. a reading equivalent to 'There is some coffee that I didn't drink'), possibly by moving covertly past negation, and therefore ending up being interpreted in an UE environment.

(5) I drank some coffee.
(6) I didn't drink some coffee. (≠ I didn't drink any coffee.)

PIs thus lie at the crossroad between language (a word is acceptable or is not acceptable in a sentence) and reasoning (a proposition follows or does not follow from another). In this study, we show that we can use

* Corresponding author.
E-mail address: milica.z.denic@gmail.com (M. Denić).

https://doi.org/10.1016/j.cognition.2021.104791
Received 9 March 2019; Received in revised form 16 May 2021; Accepted 27 May 2021
Available online 16 July 2021

the former ability, the presence of a PI in a sentence, to influence the latter, inferences made from that sentence. The importance of this result will be twofold. First, it will provide insight into our reasoning capacity. As we will see in the discussion, theories of reasoning, when combined with an appropriate theory of PI licensing, should predict this new language induced reasoning biases. Second, and as a result, they will help select between theories of PI licensing, that is, theories concerned with how language processing requires access to reasoning facts. Indeed, two classes of linguistic theories may be distinguished.

First, some theories of NPIs predict a close link between NPI licensing and checking that the corresponding monotonicity entailment holds. We will call this family of theories the *scalar theories* of NPIs.

Scalar theories of NPI licensing (Chierchia, 2006, 2013; Kadmon & Landman, 1993; Krifka, 1995; Lahiri, 1998) propose that NPIs widen the domain of quantification. For instance, while *a book* involves existential quantification over typical books, *any book* involves existential quantification over a wider domain of books, containing both typical and atypical books. In these accounts, an NPI is acceptable if it makes a point, that is, if it strengthens the meaning of a sentence. Crucially then, NPI licensing can be thought of as checking that the sentence with an NPI is logically stronger than the corresponding sentence without the NPI. In other words, it involves checking that the inference from supersets to subsets is supported, and that the inference in the other direction is not supported.

Second, some theories consider the connection between NPI licensing and inferences more loose, and as such do not by themselves predict that NPI licensing involves checking that the corresponding monotonicity entailment holds. These theories may fall in three groups.

1. *Non-veridicality theory:* A major alternative proposal for what explains NPI licensing is non-veridicality and not downward monotonicity (Giannakidou, 1998; Zwarts, 1995). Non-veridical operators are those which fail to entail the truth of their complements. For instance, negation is a non-veridical operator because *not* ϕ, for ϕ a sentence, does not entail ϕ. According to non-veridicality theory, licensing of NPIs does not need to incorporate computing the monotonicity of an environment in which an NPI occurs.
2. *Syntactic theory:* Another approach to NPI licensing that has been defended in the literature is that NPI licensing involves a syntactic dependency between an NPI and certain operators, much like subject-verb agreement dependency (Guerzoni, 2006; Herburger & Mauck, 2007; Progovac, 2000). This approach leaves room for the possibility that in resolving NPI licensing, one does not need to check which entailments are valid in the environment in which the NPI appears, but rather only needs to check that a syntactic dependency has been created between an NPI and a relevant operator.
3. *Scope theory:* Barker (2018) has recently argued that a major communicative function NPIs perform is to mark scope relations between the NPI and other operators in the sentence: simplifying somewhat, this would mean that the NPIs safe-guard language users against certain ambiguities. Again, according to this view, there is in principle no need to check entailment properties of environments in which the NPIs occur.

Similar divisions can be made between theories of PPI licensing. A version of scalar theory for NPIs exists for PPIs as well, with monotonicity properties of the environment playing an essential role in the licensing of PPIs (Nicolae, 2017). According to this theory, in order to know whether a PPI is licensed or not, one would need to compute monotonicity properties of the environment. Similarly, there are theories according to which PPI licensing is a form of syntactic dependency (Szabolcsi, 2004), or a form of scope consideration (Denić, 2015), which do not predict that processing PPIs involves computing the monotonicity properties of the environment.

In the experiments reported below, we find an effect of PIs on monotonicity inferences in various cases. Thus, we demonstrate that the presence or absence of PIs in a sentence influences which inferences subjects are willing to make, thereby demonstrating (i) that high level reasoning tasks can be influenced by what otherwise looks like innocent linguistic decorations, and (ii) that processing PIs involves monotonicity computations. From a linguistic perspective, these results are a priori coherent with the first family of theories of PI licensing; we will explain their role in psychological theories of reasoning in the discussion.

2. Previous results

Psycholinguistic studies have investigated the licensing of PIs using a variety of tasks including acceptability judgments (e.g., Drenhaus, Saddy, & Frisch, 2005; Muller & Phillips, 2018), ERP measures (e.g., Drenhaus, Graben, Saddy, & Frisch, 2006; Drenhaus, Joanna, & Julianne, 2007; Saddy, Drenhaus, & Frisch, 2004; Shao & Neville, 1998; Steinhauer, Drury, Portner, Walenski, & Ullman, 2010; Xiang, Dillon, & Phillips, 2009; Yanilmaz & Drury, 2018; Yurchenko et al., 2013), self-paced reading (e.g., Parker & Phillips, 2016; Xiang, Grove, & Giannakidou, 2013) and eye-tracking (e.g., Vasishth, Brussow, Lewis, & Drenhaus, 2008). Here we focus on two studies which jointly investigated the licensing of PIs and inferential judgments of monotonicity: Chemla, Homer, and Rothschild (2011) and Szabolcsi, Bott, and McElree (2008).

Chemla et al. (2011) collected from a group of people both upward/downward inferential judgments and NPI acceptability judgments: it was found that the inferences a particular person considers valid in a given linguistic environment predict how acceptable they would find an NPI in that same environment. This study thus provided empirical confirmation of the relationship between monotonicity properties and NPI acceptability. In fact, these results also suggest that *subjective* individual judgments of inferential properties are a better indicator of PI acceptability than objective, logical UE-ness and DE-ness. As in the experiments reported below, the study did not test all-or-nothing judgments of either NPI acceptability or monotonicity inferences, but rather looked at graded judgments. The generalization reached about the determinants of NPI acceptability were more 'graded' than those in the syntax/semantics literature. In particular, they found that DE-ness and UE-ness *together* were a better predictor of NPI acceptability than either alone was, NPIs are thus good in environments to the extent that those environments are perceived as DE and/or as not-UE.

The other relevant study on the connection between PI acceptability in an environment and the monotonicity properties of that environment is Szabolcsi et al. (2008). They report a set of experiments well designed to prompt a potential facilitation effect of the presence of an NPI on corresponding monotonicity inferences. They report on both explicit and implicit measures of inference facilitation (mere accuracy in inferential tasks, as well as reading times of phrases that presupposed the conclusion of a downward inference). They report no facilitation effect of the NPI.

In the experiments below we take another look at the question of whether PIs affect monotonicity judgments. Contrary to Szabolcsi et al. (2008), we show that PIs do in fact influence judgments of monotonicity inferences, just that these effects are (1) most noticeable in cases in which the inferential patterns are less clear to subjects (that is, not in the most basic simple UE or DE environments); (2) they are not present for all PIs in all tested configurations.

The experimental material, data, the R script used for analysis, as well as the document with the output of all of the models reported in the paper can be found athttps://github.com/milicaden/polarity-items-monotonicity-inferences.

3. Experiment 1: PIs affect the perception of monotonicity

Some environments give rise to clear (and correct) judgments of monotonicity: it is quite easy to see that 'John read a novel' entails that 'John read a book'. In such cases, adding a PI may not make the inferences any clearer, or lead people to change their mind in any way

about what inferences are supported by the environment. In this experiment, we thus looked for an effect on inferences of PIs in contexts in which the inferential patterns are less clear. *Non-monotonic* (NM) environments do not support either subset to superset or superset to subset inferences (cf. (7); neither (7a) entails (7b), nor (7b) entails (7a)). However, it has previously been shown that monotonicity judgments of these environments could be more graded, with participants reporting to a non-negligeable extent some monotonicity in one direction or another (see Chemla et al., 2011). Given this level of uncertainty as to whether these environments support upward or downward inferences, the presence or absence of a PI may then have more room to influence the judgment. Importantly for our purposes, both PPIs and NPIs are known to be acceptable at least to a certain extent in these environments: both (8a) and (8b) can be interpreted as (7a) (there is however some individual variation in terms of NPI acceptability in NM environments, cf. Rothschild, 2006, Crnič, 2014, Chemla et al., 2011, Denić, Chemla, & Tieu, 2018).

(7) a. Exactly 12 aliens saw birds.

b. Exactly 12 aliens saw doves.

(8) a. Exactly 12 aliens saw some birds.

b. Exactly 12 aliens saw any birds.

There is however an important difference between NPIs and PPIs in NM environments: PPIs like *some* can take an exceptional wide scope. For instance, (9a) has an interpretation according to which *some* takes the widest scope in the sentence (this interpretation is paraphrased in (9b)). Under the wide scope interpretation, *some doves* is no longer in a NM, but rather in a UE environment.

(9) a. Exactly 12 aliens saw some doves.

b. Some doves are such that exactly 12 aliens saw them.

This means that certain effects of PPIs like *some* on monotonicity inferences might stem not from the fact that these expressions are PPIs, but from the fact they can take wide scope. This caveat is to be kept in mind when we interpret the results of this and subsequent experiments. We will come back to it in the discussion, when we compare the results obtained with NPIs to those obtained with PPIs.

3.1. Method

3.1.1. Instructions and task

At the beginning of the task, the participants read the following instructions:

(10) *You will see pairs of sentences about aliens, who just spent last week on Earth. Imagine that you hear the first sentence, and indicate whether you would then naturally conclude that the second sentence is true.*

They were then given three examples of such pairs, call them premise-conclusion pairs. In one pair, the conclusion clearly followed from the premise (11), in a second one the conclusion clearly did not follow from the premise (12), and the third case was less clear (13).

(11) 'Each alien received a high score in all human IQ tests.' → Aliens are very intelligent.
(12) 'Few aliens visited Paris.' → All aliens visited the Eiffel Tower.
(13) 'Pink aliens have scary teeth.' → Pink aliens are the most terrifying.

The participants were instructed to record their responses on a continuous scale presented in the form of a bar by filling a portion of it red. They were told that they could use the flexibility of the red bar to report intermediate judgments, and that they would get used to it naturally. The dependent measure was the percentage of the bar filled in red. This measure will be referred to as the 'rating' given to an inference.

3.1.2. Material

The material was made of pairs of sentences, which were intended to serve as the premise and the conclusion in an inferential judgment task. These pairs of sentences were constructed from the recombination of more atomic building blocks. Crucially, among these pairs there were both valid and invalid upward and downward inferences, with and without PIs.

The building blocks used to create these inferences were as follows. First, we created a list of 8 environments: 3 UE environments (positive, Every, Many), 3 DE environments (negative, No, Few) and 2 NM environments (Exactly 12, Only 12). Second, we created a list of 12 pairs of (superset, subset) verb phrases (VPs) that could host a PI (e.g. see <PI> birds, see <PI> doves). We combined these two building blocks, environments and pairs of VPs, to obtain pairs of sentences for our inferential stimuli. Both orders of the pairs were used, i.e. superset/subset and subset/superset. Note that only superset/subset order provides a valid inference in DE environments, only subset/superset order provides a valid inference in UE environments, and neither of the orders provides a valid inference in NM environments.

Finally, for each of these pairs of sentences, we created items for which, in the premise, there was (i) no PI (for all environments), (ii) an NPI for DE and NM environments, (iii) a PPI for UE and NM environments. These possibilities correspond to all possibilities that may not be outrageously unacceptable (see discussions about the marginal acceptability of some PIs in NM environments in Rothschild, 2006, Crnič, 2014, and a quantitative evaluation in Chemla et al., 2011 and Denić et al., 2018).

Overall, we obtained 2 [superset/subset vs. subset/superset] × 12 [VPs] × (3 [UE] × 2 [PPI vs. no PI] + 3 [DE] × 2 [NPI vs. no PI] + 2 [NM] × 3 [NPI vs. PPI vs. no PI]) = 432 inference pairs. One example pair for each of the 8 environments is provided in (14)–(21).

(14) Condition: UE-positive, superset → subset, (PPI)

a. The purple alien saw (some) birds.
b. The purple alien saw doves.

(15) Condition: UE-every, superset → subset, (PPI)

a. Every alien saw (some) birds.
b. Every alien saw doves.

(16) Condition: UE-many, superset → subset, (PPI)

a. Many aliens saw (some) birds.
b. Many aliens saw doves.

(17) Condition: DE-negative, superset → subset, (NPI)

a. The purple alien didn't see (any) birds.
b. The purple alien didn't see doves.

(18) Condition: DE-no, superset → subset, (NPI)

a. No alien saw (any) birds.
b. No alien saw doves.

(19) Condition: DE-few, superset → subset, (NPI)

a. Few aliens saw (any) birds.
b. Few aliens saw doves.

(20) Condition: NM-exactly 12, superset → subset, (PPI/NPI)

a. Exactly 12 aliens saw (some/any) birds.
b. Exactly 12 aliens saw doves.

(21) Condition: NM-only 12, superset → subset, (PPI/NPI)

a. Only 12 aliens saw (some/any) birds.
b. Only 12 aliens saw doves.

These 432 items were distributed in three groups of 144 items each, so that: (a) all 12 VPs would appear in a group, (b) four different VPs were used in items which had a PPI in the premise, four different VPs were used in items which had an NPI in the premise, and four different VPs were used in items which had no PI in the premise, (c) across groups, all 12 VPs would appear with the three types of items (an NPI in the premise, a PPI in the premise, no PI in the premise). Hence, in each group there were 4 [items with different VPs] × 2 [superset/subset vs. subset/superset] × (3 [UE] × 2 [PPI vs. no PI] + 3 [DE] × 2 [NPI vs. no PI] + 2 [NM] × 3 [NPI vs. PI vs. no PI]) = 144 items. Participants were administered one of these groups of items, presented each time in a random order.

Apart from the 144 target items, the participants in each group also had to provide responses to the three training items which were administered at the beginning of the task. They were identical to the examples discussed in the instructions, and their purpose was to let participants get used to the setting and to the task.

3.1.3. Participants and exclusion criteria

75 participants were recruited through Amazon Mechanical Turk (38 females). As a result of the following two exclusion criteria, the responses of 66 participants were kept for the analysis (32 females). First, the results of one participant were excluded for them reporting not being a native speaker of English. Second, the results of eight more participants were excluded for not judging downward inferences higher in DE than in UE environments, or for not judging upward inferences higher in UE than in DE environments. The rationale for this second exclusion criterion is that, as it is likely that these judgments should be straightforward and maximally polarized, these eight participants did not understand the task in the way we expected (or were responding at random). The same exclusion criteria were applied in all four experiments reported in this paper.

3.2. Results summary

Responses given in less than 1.4 s (1% of the data) or more than 10s (9% of the data) were removed from the analysis. These numbers were chosen by a visual inspection of the distribution of RTs, with the goal of removing clear outliers. We excluded more responses falling on the slow part of the spectrum, to exclude non-spontaneous responses. This criterion was thus chosen by hand, looking only at the RTs and not the condition and responses they corresponded to. It was then copied without change for the following experiments, which provide replications of these results.

The three training items were answered as expected, with average ratings of 93% for the clearly valid inference (11), 9% for the clearly not valid one (12), and 68% for the intermediate one (13).

The left hand side of Fig. 1 summarizes the results from Experiment 1. Fig. 1 represents on the y-axis ratings of inferences with subset in premise and superset in conclusion. These inferences were valid for UE environments. These ratings thus measure the perceived UE-ness of the environment, and we refer to them as UE-ratings. On the x-axis, the ratings correspond to superset to subset inferences, which were valid in DE environments, and are accordingly referred to as DE-ratings. The graph then reports the mean ratings across participants and across the three types of environments (UE, DE, and NM). The graph shows that participants were behaving properly on these broad distinctions: disregarding the effect of PIs for the time being, UE environments ended up in the top left corner of the graph with high UE-ratings ($M = 89.7\%, SD = 13.03\%$) and low DE-ratings ($M = 29.6\ \%$, $SD = 16.2\%$), DE environments ended up in the bottom right corner of the graph with high DE-ratings ($M = 81.2\ \%$, $SD = 15.2\%$) and low UE-ratings ($M = 31.9\ \%$, $SD = 19.9\%$), and NM environments ended up in the bottom left corner of the graph with low DE and UE-ratings, even if slightly less sharply (respectively, $M = 27.7\ \%$, $SD = 17.5\%$; $M = 44.7\ \%$, $SD = 26.3\%$). We provide more detail on overall DE and UE-ratings (independent of PIs) of different environments for this and subsequent experiments in Appendix A.

The results are further separated depending on whether the premise contained a PPI, an NPI, or no PI, which is the core manipulation of interest: **does the presence of these items influence UE and DE-ratings?** In order to answer this question, we entered ratings in a model by first transforming the ratings so that they would receive a unique directional interpretation: UE-ratings were kept untransformed, but DE-ratings were reversed (x would become $100\ \% - x$). This transformed measure aligns the ratings across conditions in the following sense: it measures to what extent upward inferences follows, and to what extent downward inferences do not follow. We will refer to these as *directional ratings*. The motivation for transforming the UE-ratings and DE-ratings into a single measure of directional ratings instead of investigating the effect of PIs at each of the two ratings separately is that this ensures that any effect of PIs on inferences that we might observe is not due to PIs introducing a *yes* or *no* response bias. For instance, if a PI introduced a *yes*-bias, the presence of a PI would lead to both higher UE-ratings and higher DE-ratings, and thus lower complements of UE-ratings and DE-ratings. Combining UE-ratings with the complement of DE-ratings thus makes sure that any effect of such a bias is 'averaged out'.

Focusing on NM environments, mean participants' directional ratings seem to be, numerically, influenced by the PIs, and in particular NPIs: while the average directional ratings without a PI and with a PPI are similar ($M = 60\%$, $SD = 11.7\%$ and $M = 59.7\%$, $SD = 12.2\%$, respectively), the presence of the NPI gave rise to lower directional ratings ($M = 55.3\%$, $SD = 11.2\%$). The pertinence of these observations were confirmed by the (planned) analyses described in the following sections.

3.3. Main analyses description

In the main analyses reported in this paper (Experiments 1–4), we subset the data to items with either the PI of interest or no PI (e.g., *NPI* vs. *no PI*) in the premise in the environments of interest (e.g., *NM* environments). We fitted Bayesian linear mixed-effects regression models to participants' directional ratings with the following predictors: PI (*present* vs. *absent*), Environment instance (corresponds to different environments of the same monotonicity; e.g., in the case of NM environments these are *Exactly 12* vs. *Only 12*) and Inference direction (*superset/subset* vs. *subset/superset*). We used the Stan modeling language (Carpenter et al., 2017) and the package *brms* (Burkner, 2017). The models included maximal random-effect structures justified by the design, allowing the predictors of interest to vary by participants and by items.[1] We used the default priors of the *brms* package: a Student's t-distribution ($\nu = 3$, $\mu = 70$ and $\sigma = 40$) for the intercept, flat priors for regression coefficients, a Student's t-distribution ($\nu = 3$, $\mu = 0$ and $\sigma = 40$) for standard deviations of random effects, and LKJ $\eta = 1$ for correlation matrices. The parameters of the prior distributions are by default estimated from the data, and may vary slightly across different models.

Four sampling chains ran for at least 8000 iterations with a warm-up

[1] Items were defined as the verb phrase (VP) of the sentence rather than as the Environment instance-VP-PI combination. This is so because Environment instance and PI were treated as fixed effects.

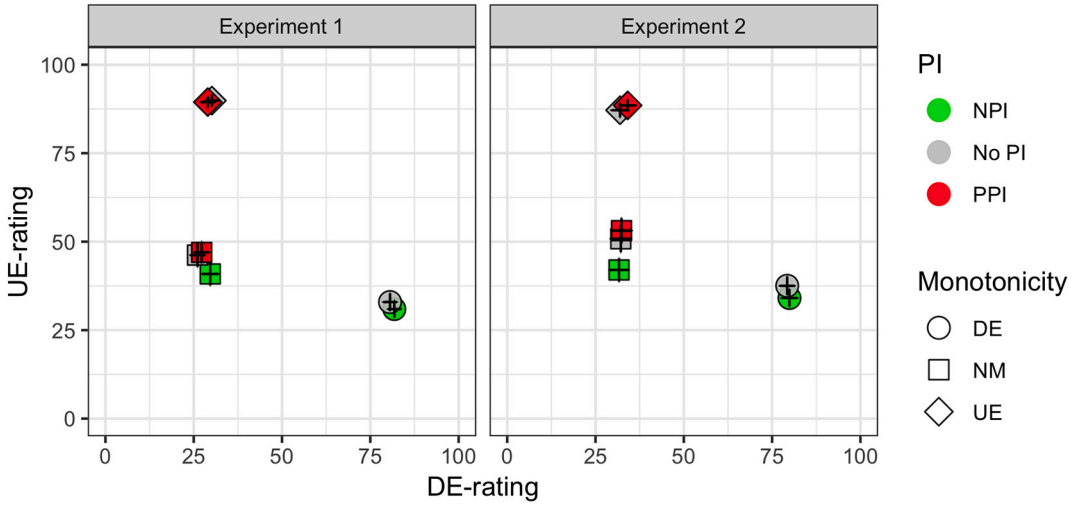

Fig. 1. Experiments 1 and 2: Mean participants' rating of the superset to subset inference (DE-rating) and of the subset to superset inference (UE-rating) in DE, UE, and NM environments depending on whether the premise contained a PPI, an NPI, or no PI. Error bars represent standard errors.

period of at least 4000 iterations for each model, resulting in at least 16,000 samples for each parameter. The exact number of iterations varied across models: for some models more iterations were necessary in order for convergence to be achieved. We assume that chains have converged when $\widehat{R}$ is below 1.1. For each parameter of interest, we report the posterior estimate $\mathbb{E}(\mu)$ and the two-side 95% credible interval (CI) based on quantiles. We also report the posterior probability that the parameter value is larger or smaller than zero ($P(\beta > 0)$ or $P(\beta < 0)$), depending on the expected direction of the effect.

As a reference point, we note that a posterior $P(\beta > 0)$ or $P(\beta < 0) > .975$ roughly corresponds to significance of a two-way frequentist inferential test with $\alpha = .05$, and $P(\beta > 0)$ or $P(\beta < 0) > .95$ to a one-way frequentist inferential test with $\alpha = .05$ (cf. Makowski, Ben-Shachar, Chen, & Lüdecke, 2019).

3.4. Analyses results

In the analyses reported here, the data was analyzed as described in Section 3.3.

The role of NPIs in NM environments We find that there is more than 99% posterior probability that the presence of NPIs (as opposed to no PIs) in the premise decreases the directional ratings in NM environments ($\mathbb{E}(\mu) = -2.18$, CI $= [-3.37, -1]$, $P(\beta < 0) = .999$).

The role of PPIs in NM environments We find that there is 55% posterior probability that the presence of PPI (as opposed to no PIs) in the premise increases the directional ratings in NM environments ($\mathbb{E}(\mu) = 0.09$, CI $= [-1.23, 1.40]$, $P(\beta > 0) = .552$).

3.5. Summary and discussion

Based on the overall pattern of results and non-contentious cases, our inference measure appears to be a relevant measure of the participants' perceived monotonicity of the environment (see also Chemla et al., 2011), according to which DE and UE environments are perceived as such, and NM environments as intermediate. The main finding is that there is strong evidence that the presence of an NPI in the premise decreases directional ratings in NM environments. In other words, the presence of NPIs makes participants perceive an NM environment as less UE and/or more DE as compared to when no PI is in the premise. No strong conclusions can be drawn about the influence of PPIs on monotonicity inferences in NM environments in Experiment 1.

Note that the main observed effect (i.e. the effect of NPIs on inferential judgments in NM environments) cannot be explained away as a regression to the mean (i.e. as more random responses than in the no PI condition leading to judgments overall closer to 50% in the PI condition). The reason why this alternative explanation is not viable is that the NPI seems to be making people perceive a NM as less UE, pushing them further away from the mean (50%) UE-rating than the baseline (no PI) condition is (cf. Fig. 1 and the analyses in Section 8).

4. Experiment 2: Replication (with PIs also in conclusions)

In Experiment 1, PIs were present only in the premises, to assess their role as a guide for 'future' inferences, but this creates an asymmetry between the premise and the conclusion, which may obscure the effect of the PI (for instance, as the PI was present in the premise but not in the conclusion, it might have been quite easy for participants to ignore it). We thus ran Experiment 2, which was identical to Experiment 1, except that whenever a PI was present in the premise, it was also present in the conclusion.

Roughly put, Experiment 2 provides us with a replication of the previous result.

4.1. Method

4.1.1. Instructions and task

Instructions and task were identical to those in Experiment 1.

4.1.2. Material

Material was identical to those used in Experiment 1 except that the conclusion sentences contained a PI whenever the premise did.

4.1.3. Participants

72 participants were recruited through Amazon Mechanical Turk (35 females). One participant was excluded from the analysis for reporting not being a native speaker of English and seven more for not showing a difference in perceived monotonicity of UE and DE environments (same exclusion criteria as in Experiment 1). 64 participants were thus kept for the analysis (28 females).

4.2. Results summary

As in Experiment 1, responses given in less than 1.4 s (6% of the data) or more than 10s (7% of the data) were removed.

Training items were answered as expected: the clearly valid inference received an average rating of 93%, the clearly not valid one 9%, the intermediate one 62%. The right hand side of Fig. 1 summarizes the

results from Experiment 2. As before, it represents on the x-axis mean participants' rating of superset to subset inferences (DE-ratings), and on the y-axis mean participants' rating of subset to superset inferences (UE-ratings), across three types of environments (UE, DE, and NM), depending on whether the premise and conclusion contained a PPI, an NPI, or no PI.

As in Experiment 1, if we first disregard the effect of PIs, the three types of environments UE, DE and NM behave distinctly, as expected. We also observe a similar pattern as before for the role of PIs in NM environments, with inferences without a PI vs. with a PPI receiving similar directional ratings ($M = 59.2\%$, $SD = 14.9\%$ and $M = 60.1\%$, $SD = 14.1\%$, respectively), and the presence of NPIs leading to lower directional ratings ($M = 54.7$ % , $SD = 12.5\%$).

4.3. Analyses results

The data were analyzed as described in Section 3.3.

The role of NPIs in NM environments We find that there is more than 99% posterior probability that the presence of NPIs (as opposed to no PIs) in the premise decreases the directional ratings in NM environments ($\mathbb{E}(\mu) = -2.25$, CI $= [-3.48, -1.02]$, $P(\beta < 0) = 0.999$).

The role of PPIs in NM environments We find that there is 84% posterior probability that the presence of PPIs (as opposed to no PIs) in the premise increases the directional ratings in NM environments ($\mathbb{E}(\mu) = 0.63$, CI $= [-0.64, 1.92]$, $P(\beta > 0) = 0.84$).

4.4. Summary and discussion

In Experiment 1, we tested whether the presence of NPIs and PPIs as compared to no PIs in the premise has an influence on the monotonicity inferences with conclusions without PIs. Experiment 2 differed from Experiment 1 only in that whenever a PI was present in the premise, it was also present in the conclusion. The results of Experiment 2 confirm those of Experiment 1 when it comes to the influence of NPIs on monotonicity inferences: the presence of NPIs makes participants perceive an NM environment as less UE and/or more DE than when it contains no PIs. As in Experiment 1, no strong conclusions can be drawn about the influence of PPIs on monotonicity inferences in NM environments in Experiment 2 either.

5. Doubly negative environments

In Experiments 1 and 2, we have established that there is an influence of PIs on monotonicity inferences in NM environments. NM environments are characterized by two aspects: first, inferential judgments in these environments are not straightforward; second both PPIs and NPIs are acceptable in these environments, at least to some extent. In the continuation of the paper, we will extend the inquiry to another type of environments with these two properties.

These are the so-called *doubly-negative* (DN) environments as in (22), which are a type of UE environments. As mentioned above, both PPIs and NPIs are known to be licit in these environments: this means, for NPIs like *any*, that they are acceptable in those environments, and for PPIs like *some*, that they can be interpreted with narrow scope under both operators; for instance, (22a) can be interpreted as (22b):

(22) a. Every alien who did not see some doves is hairy.

b. Every alien who did not see any doves is hairy.

Importantly, due to the combination of two DE operators (*n't* is DE and *every* is DE in its restrictor), the NPI *any* ends up appearing in a global UE environment in (22b). This example thus shows that global logical properties cannot (always) be responsible for NPI licensing. Therefore researchers who advocate a monotonicity-based approach to licensing are led to go local, i.e. propose that the system that checks the acceptability of a given PI in a sentence S has access to constituents of S, and that PIs are licensed if at least one of the constituents of S they are in has the appropriate monotonicity properties (Gajewski, 2005; Homer, 2020). For concreteness, in a sentence like (22b), this system can single out the VP of the relative clause and compute its monotonicity with respect to the position of the NPI *any* (monotonicity is a property of functions; to evaluate what we loosely call the monotonicity of a constituent, one has to abstract over a position within this constituent, e.g., the position of the PI): this constituent turns out to be DE w.r.t. this position. As the licensing condition just requires that a PI be in at least one constituent which has the appropriate monotonicity w.r.t. its position, the NPI *any* is licensed in (22b).

There is, however, yet another interesting possibility for why NPIs are acceptable in DN environments. It is possible that monotonicity inferences are so hard in these environments that people wrongly consider them DE to some extent. The NPIs would thus be licensed in these environments because of the subjective (wrong) perception of their monotonicity. There are thus two distinct options for how NPIs may influence monotonicity inferences in DN environments: they may lead to the local environment being perceived (correctly) as more DE/less UE, which would improve the perception of global environment as UE and not DE. Alternatively, they may influence directly the perception of monotonicity properties of global environment, leading to it being perceived (incorrectly) as more DE/less UE. In addition to documenting the effect of PIs on monotonicity in a new type of environment, this inquiry could thus also be directly informative about the status of NPI licensing in DN environments (are they licensed because the local environment is DE, or because the global environment is wrongly perceived as DE/not UE?). We will discuss this later on, as it will be easier to do so with the results in place.

6. Experiment 3: Doubly negative is not positive

Experiment 3 tested the effect of the PIs on monotonicity inferences in environments with two accumulating DE operators, such as (22).

6.1. Method

6.1.1. Instructions and task

Instructions and task were identical to those in Experiment 1 and 2.

6.1.2. Material

The stimuli were identical to those used in Experiment 1, except for the addition of two DN environments. These were presented with a PPI, an NPI, or no PI in the premise (and no PI in the conclusion). We thus obtained 2 [superset/subset vs. subset/superset] × 12 [VPs] × (3 [UE] × 2 [PPI vs. no PI] + 3 [DE] × 2 [NPI vs. no PI] + 2 [NM] × 3 [NPI vs. PPI vs. no PI] + 2 [DN] × 3 [NPI vs. PPI vs. no PI]) = 576 inference pairs. An example of premise-conclusion pair for each of the two DN environments is in (23) and (24). These 576 items were split into three groups with 192 items, which satisfied the same conditions as the groups in Experiment 1. Participants were randomly administered to one of the three groups.

(23) Condition: DN-Every-not, subset → superset, (PPI/NPI)

a. Every alien who did not see (some/any) doves is hairy.
b. Every alien who did not see birds is hairy.

(24) Condition: DN-No-without, subset → superset, (PPI/NPI)

a. No alien spent a year without seeing (some/any) doves.
b. No alien spent a year without seeing birds.

6.1.3. Participants

112 participants were recruited through Amazon Mechanical Turk (69 females). Seven participants were excluded from the analysis for

reporting not being a native speaker of English and 13 more for not showing much difference in perceived monotonicity of UE and DE environments (same exclusion criteria as in Experiments 1 and 2). 92 participants were thus kept for the analysis (53 females).

6.2. Results summary

As in Experiments 1 and 2, responses given in less than 1.4 s (4% of the data) or more than 10s (13% of the data) were removed from the analysis. Training items were answered as expected: the clearly valid inference received an average rating of 92%, the clearly not valid one 8.3%, the intermediate one 67%.

The left hand side of Fig. 2 summarizes the results from Experiment 3. Fig. 2 represents on the x-axis mean participants' rating of superset to subset inference (DE-rating), and on the y-axis mean participants' rating of subset to superset inference (UE-rating), across four types of environments (UE, DE, NM, and DN), depending on whether the premise contained a PPI, an NPI, or no PI.

Disregarding whether and which PI was present in the premise, the four types of environments (UE, DE, NM and DN) are well-separated and they show up where they could have been expected. DN environments are quite interesting in this respect: as a reminder, DN environments are in fact plain UE environments. Nonetheless, they seem to behave in a more intermediate fashion, and they are much closer to NM than to UE environments.

Looking first at the replication of the results from Experiments 1 and 2 in NM environments, mean participants' directional ratings were (i) $M = 55.9\%$ ($SD = 13.4\%$) when NPI is in the premise, (ii) $M = 57.9\%$ ($SD = 14\%$) when PPI is in the premise, and (iii) $M = 57.1\%$ ($SD = 13.3\%$) when no PI is in the premise.

Moving to DN environments, mean participants' directional ratings were (i) $M = 53.7\%$ ($SD = 16.8\%$) when NPI is in the premise, (ii) $M = 61.8\%$ ($SD = 14.4\%$) when PPI is in the premise in, and (iii) $M = 56.9\%$ ($SD = 14.7\%$) when no PI is in the premise.

6.3. Analyses results

The data were analyzed as described in Section 3.3.

The role of NPIs in NM environments We find that there is more than 95% posterior probability that the presence of NPIs (as opposed to no PIs) decreases the directional ratings in NM environments ($\mathbb{E}(\mu) = -1.11$, CI = [−2.35,0.11], $P(\beta < 0) = .965$).

The role of PPIs in NM environments We find that there is 88% posterior probability that the presence of PPIs (as opposed to no PIs) in the premise increases the directional ratings in NM environments ($\mathbb{E}(\mu) = 0.59$, CI = [−0.42,1.59], $P(\beta > 0) = .88$).

The role of NPIs in DN environments We find that there is 89% posterior probability that the presence of NPIs (as opposed to no PIs) in the premise decreases the directional ratings in DN environments ($\mathbb{E}(\mu) = -0.85$, CI = [−2.27,0.59], $P(\beta < 0) = .887$).

The role of PPIs in DN environments We find that there is more than 99% posterior probability that the presence of PPIs (as opposed to no PIs) in the premise increases the directional ratings in DN environments ($\mathbb{E}(\mu) = 2.06$, CI = [0.77,3.38], $P(\beta > 0) = .998$).

6.4. Summary and discussion

In Experiment 3, we again find strong evidence — albeit somewhat less so than in Experiments 1 and 2 — for the influence of NPIs on monotonicity inferences in NM environments: the presence of NPIs makes participants perceive an NM environment as less UE and/or more DE than when it contains no PIs. As in Experiments 1 and 2, we are not in a position to draw strong conclusions about the influence of PPIs on monotonicity inferences in NM environments.

In addition to replicating the results from Experiments 1 and 2, in Experiment 3 we tested whether the presence of NPIs and PPIs as compared to no PIs in DN environments has an influence on global monotonicity inferences. While we cannot draw strong conclusions about the influence of NPIs on monotonicity inferences in DN environments, an effect of PPIs on the global monotonicity inferences was found in DN environments, whereby there is strong evidence that the presence of PPIs makes participants perceive DN environments as more UE and/or less DE (recall that while these environments are in fact UE, they are not judged as such to the same extent as simple UE environments).

7. Experiment 4: Doubly negatives, a replication (with a different arrangement of the items)

Experiment 4 is a conceptual replication of the previous experiments. The main difference between Experiment 4 and previous experiments is that, unlike in the previous experiments, the presence of a PI in a given environment instance (i.e. specific instances of 3 UE, 3 DE, 2 NM, or 2 DN environments) was a between-participants factor.[2] This change of setting decreases noise in certain respects (there can be no spill-over effect of PIs in the same environment), but increases noise in other respects (we are looking at the effect of PIs across different groups of participants, which may have different baseline judgments). Importantly, most of our results replicate in this setting.

7.1. Method

7.1.1. Instructions and task

Instructions and task were identical to those in Experiments 1, 2 and 3.

7.1.2. Material

Materials were the same as those used in Experiment 3, with two differences. First, the total number of items was reduced to 480 by reducing the number of different VPs from 12 to 10. Second, the participants were split into four groups (instead of three) in such a way that each participant sees a given environment instance either with an NPI, or with a PPI, or without a PI. Because of this, there were 2 [superset/subset vs. subset/superset] × 10 [VPs] × (3 [UE] + 3 [DE] + 2 [NM] + 2 [DN]) = 200 items per group.

7.1.3. Participants

81 participants were recruited through Amazon Mechanical Turk (43 females). Four participants were excluded from the analysis for reporting not being a native speaker of English and six more for not showing much difference in perceived monotonicity of UE and DE environments (same exclusion criteria as in Experiments 1–3). 71 participants were thus kept for the analysis (36 females).

7.2. Results summary

As in Experiments 1–3, we removed responses given in less than 1.4 s (8% of the data) or more than 10s (10% of the data). Training items were answered as expected: the clearly valid inference received an average rating of 89.5%, the clearly not valid one 8.2%, the intermediate one 60.7%.

The right hand side of Fig. 2 summarizes the results of Experiment 4. Fig. 2 represents on the x-axis mean participants' rating of superset to subset inference (DE-rating), and on the y-axis mean participants' rating of subset to superset inference (UE-rating), across four types of environments (UE, DE, NM, and DN), depending on whether the premise

[2] This means that for a majority of participants in Experiment 4 we do not have data for both a PI of interest (NPI or PPI) and a baseline (no PI) in an environment of interest (NM or DN). Because of this, the maximal random effect structure justified by the design of Experiment 4 for the analyses of interest doesn't include random by-participant slopes for PI.

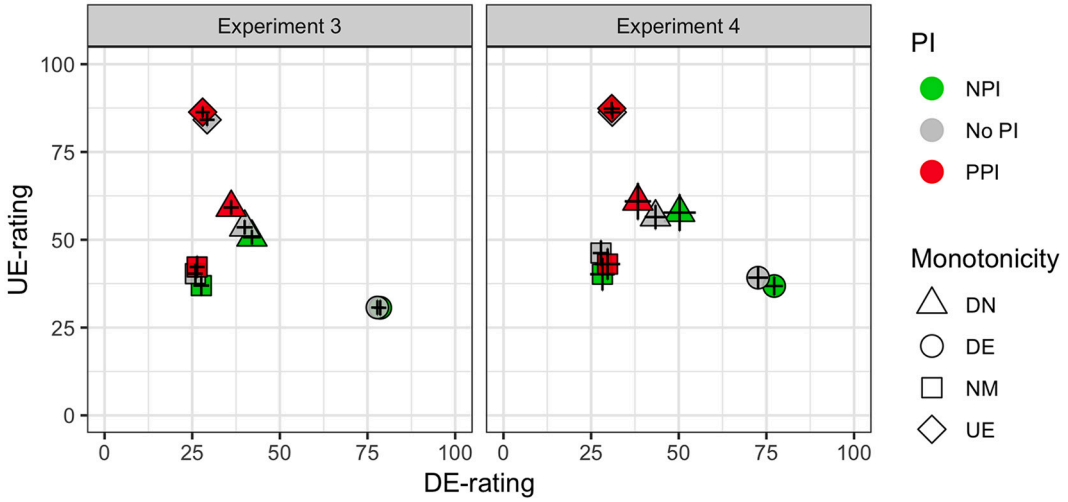

Fig. 2. Experiments 3 and 4: Mean participants' rating of the superset to subset inference (DE-rating) and of the subset to superset inference (UE-rating) in DE, UE, NM, and DN environments depending on whether the premise contained a PPI, an NPI, or no PI. Error bars represent standard errors.

contained a PPI, an NPI, or no PI.

Disregarding for the time being the role of the PIs on inferential judgments, the four environments (DE, UE, NM, DN) are well separated: these results seem to be qualitatively identical to the results in Experiment 3.

Looking again first at the effect of PIs in NM environments, mean participants' directional ratings are (i) $M = 56.3\%$ ($SD = 11.6\%$) when NPI is in the premise, (ii) $M = 56.4\%$ ($SD = 11.1\%$) when PPI is in the premise in, and (iii) $M = 59.3\%$ ($SD = 15.3\%$) when no PI is in the premise.

In DN environments, mean participants' directional ratings were (i) $M = 52.6\%$ ($SD = 25.2\%$) when NPI is in the premise, (ii) $M = 61.6\%$ ($SD = 17.4\%$) when PPI is in the premise in, and (iii) $M = 56.3\%$ ($SD = 22.4\%$) when no PI is in the premise.

7.3. Analyses results

The data were analyzed as described in Section 3.3.

The role of NPIs in NM environments We find that there is 98% posterior probability that the presence of NPIs (as opposed to no PIs) decreases the directional ratings in NM environments ($\mathbb{E}(\mu) = -1.64$, CI $= [-3.08, -0.19]$, $P(\beta < 0) = .984$).

The role of PPIs in NM environments We find that there is 10% posterior probability that the presence of PPIs (as opposed to no PIs) in the premise increases the directional ratings in NM environments ($\mathbb{E}(\mu) = -1$, CI $= [-2.54, 0.55]$, $P(\beta > 0) = .104$).

The role of NPIs in DN environments We find that there is more than 90% posterior probability that the presence of NPIs (as opposed to no PIs) in the premise decreases the directional ratings in DN environments ($\mathbb{E}(\mu) = -2.72$, CI $= [-6.45, 1.09]$, $P(\beta < 0) = .924$).

The role of PPIs in DN environments We find that there is 84% posterior probability that the presence of PPIs (as opposed to no PIs) in the premise increases the directional ratings in DN environments ($\mathbb{E}(\mu) = 1.92$, CI $= [-1.90, 5.66]$, $P(\beta > 0) = .838$).

7.4. Summary and discussion

In Experiment 4, we replicate the findings of previous experiments related to the influence of NPIs on monotonicity inferences in NM environments. We again do not draw strong conclusions about an influence of PPIs on monotonicity inferences in NM environments. In Experiment 4, we are not in a position to draw strong conclusions about an influence of NPIs or PPIs on monotonicity inferences in DN environments either.

8. Combined analyses

With the results of the four experiments in place, we conducted three sets of combined analyses over the four experiments. The first set of combined analyses consists of meta-analyses over the four experiments investigating the influence of PIs in different environments. The second set of combined analyses investigates which of the two dimensions of inferences (upward and downward) PIs have an influence on in different environments. Finally, the third set of combined analyses investigates potential differences between environments in terms of how PIs influence inferences.

We conducted two additional sets of combined analyses which will not be central to the following discussion, and are thus reported in an Appendix. Appendix A analyzes how participants' DE-ratings and UE-ratings vary across four environments (DE, UE, NM, DN) studied in our experiments, independently of PIs. Appendix B investigates whether there are differences between the two instances of NM environments and between the two instances of DN environments from our experiments in terms of how PIs influence monotonicity inferences.

8.1. Meta-analyses

The data were analyzed as described in Section 3.3.

The role of NPIs in NM environments We find that there is more than 99% posterior probability that the presence of NPIs in the premise (as opposed to no PIs) decreases the directional ratings in NM environments when pooling the results of all experiments ($\mathbb{E}(\mu) = -1.73$, CI $= [-2.34, -1.13]$, $P(\beta < 0) = .999$).

The role of PPIs in NM environments We find that there is 82% posterior probability that the presence of PPIs in the premise (as opposed to no PIs) increases the directional ratings in NM environments when pooling the results of all experiments ($\mathbb{E}(\mu) = 0.26$, CI $= [-0.32, 0.86]$, $P(\beta > 0) = .817$).

The role of NPIs in DN environments We find that there is more than 95% posterior probability that the presence of NPIs in the premise (as opposed to no PIs) decreases the directional ratings in DN environments when pooling the results of all experiments ($\mathbb{E}(\mu) = -1.13$, CI $= [-2.42, 0.15]$, $P(\beta < 0) = .96$).

The role of PPIs in DN environments We find that there is more than 99% posterior probability that that the presence of PPIs in the premise (as opposed to no PIs) increases the directional ratings in DN environments when pooling the results of all experiments ($\mathbb{E}(\mu) = 2.02$, CI $= [0.88, 3.16]$, $P(\beta > 0) = .999$).

These meta-analysis confirm the effect of NPIs on monotonicity

inferences in NM environments in that NPIs make participants perceive NM environments as less UE and/or more DE, in line with the results of all four experiments. Interestingly, these analyses strongly suggest that NPIs have such an effect in DN environments as well.

As for PPIs, neither the results of the four experiments nor the present meta-analysis allow to draw strong conclusions about their influence on monotonicity inferences in NM environments. The present meta-analysis however demonstrates that PPIs influence monotonicity inferences in DN environments in that they make participants perceive DN environments as more UE/less DE, in line with the results of Experiment 3.

8.2. Influence of PIs on two inferential dimensions

Across the statistical analyses reported separately for the four experiments, as well as for those in Section 8.1, the dependent measure was that of directional ratings, as motivated in Section 3.2.

For completeness, in the following analyses, reported in Table 1, we pool the data from the four experiments to investigate the influence of PIs on monotonicity inferences separately for two inferential directions: *superset/subset* (i.e. DE-ratings) or *subset/superset* (i.e. UE-ratings). To this end, we subset the data to items with either the PI of interest or no PI (e.g., *NPI* vs. *no PI*) in the premise in an environment of interest (e.g., *NM*) with inference direction of interest (e.g. *superset/subset*). We fitted Bayesian linear mixed-effects regression models to participants' responses with the following predictors: PI (*present* vs. *absent*), Environment instance (e.g., in the case of NM environments these are *Exactly 12* vs. *Only 12*), with the maximal random-effect structures justified by the design. We used the default priors of the *brms* package (cf. Section 3.3).

Four sampling chains ran for at least 8000 iterations with a warm-up period of at least 4000 iterations for each model, resulting in at least 16,000 samples for each parameter. The exact number of iterations varied across models: for some models more iterations were necessary in order for convergence to be achieved. See Section 3.3 for details on results reporting and interpretation.

There is strong evidence that NPIs decrease UE-ratings and increase DE-ratings in NM environments. Similarly, there is strong evidence that PPIs increase UE-ratings and decrease DE-ratings in DN environments. Finally, there is some evidence that NPIs increase DE-ratings in DN environments, and that PPIs increase UE-ratings in NM environments. No strong conclusions can be drawn about the influence of NPIs on UE-ratings in DN environments, or about the influence of PPIs on DE-ratings in NM environments.

Note however that these secondary analyses carry less weight than the analyses on directional ratings: any effect found in the analyses that separate DE-ratings from UE-ratings could potentially be due to PIs introducing a *yes*- or *no*-response bias, as discussed in Section 3.2. These analyses should be thus only taken as suggestive for how PIs influence DE-ratings and UE-ratings separately.

Table 1
Influence of NPIs and PPIs on DE and UE ratings in NM and DN environments.

	Influence of PI on UE-rating	Influence of PI on DE-rating
NPIs in NM	NPI: $M = 39.7$ % , $SD = 26.3$% no PI: $M = 45.4$ % , $SD = 28.6$% $\mathbb{E}(\mu) = -2.54$, CI = $[-3.51, -1.59]$ $P(\beta < 0) = 1$	NPI: $M = 29.2$ % , $SD = 20.7$% no PI: $M = 27.8$ % , $SD = 19.9$% $\mathbb{E}(\mu) = 0.93$, CI = [0.07,1.81] $P(\beta > 0) = .982$
NPIs in DN	NPI: $M = 52.7$ % , $SD = 24.1$% no PI: $M = 54.9$ % , $SD = 24$% $\mathbb{E}(\mu) = -1.15$, CI = [−2.98,0.63] $P(\beta < 0) = .903$	NPI: $M = 44.39$ % , $SD = 25.5$% no PI: $M = 41.4$ % , $SD = 23.7$% $\mathbb{E}(\mu) = 1.39$, CI = [−0.22,3] $P(\beta > 0) = .956$
PPIs in NM	PPI: $M = 46.2$ % , $SD = 27.9$% no PI: $M = 45.4$ % , $SD = 28.6$% $\mathbb{E}(\mu) = 0.92$, CI = [−0.05,1.89] $P(\beta > 0) = .97$	PPI: $M = 28.6$ % , $SD = 20.6$% no PI: $M = 27.8$ % , $SD = 19.9$% $\mathbb{E}(\mu) = 0.24$, CI = [−0.40,0.88] $P(\beta < 0) = .227$
PPIs in DN	PPI: $M = 59.6$ % , $SD = 22.6$% no PI: $M = 54.9$ % , $SD = 24$% $\mathbb{E}(\mu) = 2.03$, CI = [0.45,3.59] $P(\beta > 0) = .993$	PPI: $M = 36.7$ % , $SD = 21$% no PI: $M = 41.4$ % , $SD = 23.7$% $\mathbb{E}(\mu) = -2.1$, CI = $[-3.74, -0.45]$ $P(\beta < 0) = .993$

8.3. PIs — Environments interactions

In the experiments reported in this paper, we have focused on the influence of PIs on inferential judgments in NM and DN environments. The motivation for this is that these environments, because of their complexity, may leave more room for PIs to influence inferences than (simple) UE or DE environments. One may wonder however whether it is indeed the case that the influence of PIs on inferences are more robust in NM or DN than in DE or UE environments. The analyses reported below aim at answering this question.

In the following analyses, we thus compare pairs of environments with respect to the influence of PIs on monotonicity inferences. To this end, we subset the data to items with either the PI of interest or no PI (e.g. *NPI* vs. *no PI*) in the premise in a pair of environments of interest (e.g. NM and DE environments). We fitted Bayesian linear mixed-effects regression models to participants directional ratings with the following predictors: PI (*present* vs. *absent*), Environment (e.g. *NM* vs. *DE*), PI-Environment interaction term, and Inference direction (*superset/subset* vs. *subset/superset*), with the maximal random-effect structures justified by the design. We used the default priors of the *brms* package (cf. Section 3.3).

Four sampling chains ran for at least 2000 iterations with a warm-up period of at least 1000 iterations for each model, resulting in at least 4000 samples for each parameter. The exact number of iterations varied across models: for some models more iterations were necessary in order for convergence to be achieved. See Section 3.3 for details on results reporting and interpretation.

NPIs in NM vs. DE environments We find that there is 98% posterior probability that the presence of NPIs in the premise decreases the directional ratings more in NM environments as opposed to DE environments ($\mathbb{E}(\mu) = -0.56$, CI = $[-1.06, -0.02]$, $P(\beta < 0) = .977$).

NPIs in DN vs. DE environments We find that there is 99% posterior probability that the presence of NPIs in the premise decreases the directional ratings more in DN as opposed to DE environments ($\mathbb{E}(\mu) = -0.79$, CI = $[-1.49, -0.08]$, $P(\beta < 0) = .987$).

PPIs in NM vs. UE environments We find that there is 48% posterior probability that the presence of PPIs in the premise increases the directional ratings more in NM as opposed to UE environments ($\mathbb{E}(\mu) = -0.01$, CI = $[-0.47, 0.43]$, $P(\beta > 0) = .475$).

PPIs in DN vs. UE environments We find that there is more than 99% posterior probability that the presence of PPIs in the premise increases the directional ratings more in DN as opposed to UE environments ($\mathbb{E}(\mu) = 1.15$, CI = [0.69,1.62], $P(\beta > 0) = .999$).

There is strong evidence that NPIs have a more visible effect on monotonicity inferences in NM and DN environments than in DE environments, and for PPIs this is the case in DN environments as opposed to UE environments. Monotonicity inferences in complex environments, such as NM and DN environments, thus seem to be more prone to being modified by cues such as PIs.

For completeness, we further report PI-environment interactions for DN and NM environments.

NPIs in DN vs. NM environments We find that there is 33% posterior probability that the presence of NPIs in the premise decreases the directional ratings more in NM as opposed to DN environments ($\mathbb{E}(\mu) = 0.14$, CI = $[-0.48, 0.76]$, $P(\beta < 0) = .329$).

PPIs in DN vs. NM environments We find that there is more than 99% posterior probability that the presence of PPIs in the premise increases the directional ratings more in DN as opposed to NM environments ($\mathbb{E}(\mu) = 1.15$, CI = $[0.52, 1.77]$, $P(\beta > 0) = .999$).

9. Discussion: PIs and environments

Let us summarize the findings: which PIs influence monotonicity inferences in which environments?

Let us start with NPIs. Across four experiments, we find an effect on NPIs on monotonicity inferences in NM environments. This is further confirmed by the meta-analysis reported in Section 8.1. This effect is two-fold: the presence of NPIs increases DE-ratings and decreases UE-ratings (cf. Section 8.2 and Table 1).

Furthermore, we find that the effect of NPIs on monotonicity inferences is more robust in NM than in DE environments (cf. Section 8.3). Where does this asymmetry between NM and DE environments come from? One possibility is that it could be well-explained by *ceiling* effects in DE environments (in our experiments, as well as in Szabolcsi et al., 2008). Indeed, for DE environments, participants may report high DE judgments and low UE judgments as much as they possibly can even without an NPI, leaving little room for an NPI to make the judgments even more extreme.

Let us now move to NPIs in DN environments. While we were not able to draw conclusions about the influence of NPIs on monotonicity inferences based on separate analyses of Experiments 3 and 4, the meta-analysis reported in Section 8.1 suggests that NPIs influence monotonicity inferences in DN environments as well. Furthermore, we do not find evidence that the influence of NPIs on monotonicity inferences differs in NM as compared to DN environments (cf. Section 8.3). Our results thus suggest that NPIs influence inferences in DN environments too.

This fact may speak to the question of why NPIs are licensed in DN environments. Note that the environments in which NPIs are acceptable are not all perceived as DE (DE-ratings in NM and DN environments are lower than in DE environments, cf. Appendix A), but all of them are perceived as not UE (UE-ratings in DE, NM, and DN environments are lower than in UE environments, cf. Appendix A). The latter is perhaps unsurprising for DE and NM environments, but it is remarkable for DN environments, which are in fact UE. If the presence of an NPI makes one perceive a DN environment as less UE and more DE, this opens the possibility that the acceptability of NPIs in such environments is due (at least in part) to the perception that the environment is, at a global level, not UE. This is a significant departure from current approaches (although see Chemla et al., 2011 for further evidence that global monotonicity inferences may play a role in NPI licensing). This perspective emerges here from the systematic collection of inferential and acceptability judgments, and it could naturally be put to further tests. This theoretical option may also illustrate a particular type of cognitive approach to linguistic generalizations in general, in which subjective and potentially 'fallacious' judgments have their say in grammar and grammatical theorizing.

Let us now move to PPIs. We do not find strong evidence that they influence monotonicity inferences in NM environments in any of the four experiments, nor in the meta-analysis reported in Section 8.1.[3]

Experiment 2 is of particular interest with respect to this. Namely, in Experiment 2, both the premise such as (25a), and the conclusion such as (25b), had a PI.

(25) a. Exactly 12 aliens saw some doves.

b. Exactly 12 aliens saw some birds.

Recall that PPIs like *some* can take wide scope (cf. Section 3), and that in such an event *some birds* is in an UE rather than in NM environment in (25a). If *some* took wide scope in both the premise and the conclusion, NM environments with PPIs should look like UE environments when it comes to monotonicity inferences. That the results of Experiment 2 are inconclusive about the effect of PPIs on monotonicity inferences in NM environments in Experiment 2 thus suggests that people tend to not assign wide scope to *some* in NM environments such as (25a). This further suggests that the fact that we do not find evidence for the effect of PPIs on inferences in NM environments across four experiments is not due to the availability of two interpretations of the premise.

How about PPIs in DN environments? Interestingly, we do find evidence for an effect of PPIs on monotonicity inferences in DN environments in Experiment 3 (albeit not in Experiment 4), as well as in the meta-analysis reported in Section 8.1. There is also evidence that PPIs have greater influence on monotonicity inferences in DN environments than in NM environments (cf. Section 8.3).

Why do we observe the effect of PPIs in DN environments but not in NM environments? This could be related to the fact that the baseline inferential judgments may leave more room for PPI effects in DN environments than in NM environments. Let us explain. While neither of the two environments are DE, DN environments are judged on average as more DE than NM environments (cf. Appendix A). This means that there is more room for improvement in terms of DE-rating in DN environments as opposed to NM environments. Accordingly, a PPI effect pushing judgments away from this mistake may be easier to observe for DN environments.

However, this result must be interpreted with care in light of the availability of the wide scope with PPI *some* (cf. Section 3).

For instance, (26a) has an interpretation according to which *some* takes the widest scope in the sentence (this interpretation is paraphrased in (27)). Note that under this interpretation, the subset to superset inference from (26a) to (26b) follows.

(26) a. Every alien who did not see some doves is hairy.

b. Every alien who did not see birds is hairy.

(27) Some doves are such that every alien who didn't see them is hairy.

This means that, if for whichever reason the wide scope of *some* is easier to obtain in DN environments than in NM environments, and if the wide scope interpretation of *some* makes it easier to see that the inference from (26a) to (26b) follow, one could explain the effect of PPI in DN environments without relating PPI licensing to monotonicity inferences.

We acknowledge that more work is needed to understand better the effect of PPIs on monotonicity inferences in DN environments together with the lack of evidence for such effect in NM environments.

10. Discussion: Three routes from PIs to inferences

In this study, we have explored whether PIs influence monotonicity inferences. Our results demonstrate that this is indeed the case. This was demonstrated most convincingly for NPIs in NM environments. What is the mechanism behind this influence? We discuss three possibilities here, relating them to the two families of approaches to NPI licensing introduced in Section 1.

10.1. PIs induce biases as priming

In the introduction, we have distinguished between two families of linguistic theories of NPI licensing: those according to which processing of NPIs involves monotonicity computation, and those according to which it does not.

What happens when sentences with NPIs are parsed, what elementary operations does this involve? According to the first family of approaches to NPI licensing introduced in Section 1 (i.e. scalar theories of NPI licensing), a successful parsing of a sentence with an NPI involves ensuring that the environment in which the NPI is found has the right

[3] The analysis conducted separately on two inferential dimensions reported in Section 8.2 provides however suggestive evidence that PPIs may influence UE-ratings, but not DE-ratings in NM environments.

monotonicity properties: DE-ness and not UE-ness. This may be done in different ways, for instance, by running mental simulations as in mental models approach to linguistic inferences (Johnson-Laird, 1983, a.o), or by proof solvers as in mental logic approach to linguistic inferences (Braine, 1978, a.o.). To the extent that the computation of monotonicity inferences recruits the same mechanisms during language parsing and in tasks down the line (such as reasoning tasks), priming effects may be expected. In other words, PIs which would have triggered the computation of monotonicity inferences at parsing time, would increase the likelihood of similar monotonicity computations being performed when one reasons with the parsed sentence.

10.2. PIs induce biases as a side effect of the meaning of PIs

We now discuss a second possibility for what mechanism may be behind the influence of PIs on reasoning. Namely, if sentences with a PI have a different semantic interpretation from sentences without a PI, the effects we observe may follow from this meaning difference.

Focusing on NPIs, to our knowledge the only approach to NPI licensing according to which a sentence such as (28) has a different semantic interpretation from the sentence (29) are the scalar theories of NPI licensing (cf. Section 1). Recall that according to this approach, NPIs induce domain widening, which means that (30) holds, and hence that {y | y saw *any* birds} may be a superset of {x | x saw birds}.

(28) No aliens saw birds.
(29) No aliens saw any birds.
(30) {x | x saw birds} ⊆ {y | y saw *any* birds}.

We will now discuss whether any theory of reasoning with quantified sentences in combination with a scalar theory of NPI licensing captures the effect of NPIs on monotonicity inferences.

Approaches to human reasoning which have focused on inferences people draw from quantified sentences can be divided into three main families: mental models approaches, mental logic approaches, and probabilistic approaches. We discuss these in turn in relation to monotonicity inferences; we take as a working example downward inference in the scope of *no* from (28) to (31), and discuss how adding an NPI to (28) as in (29) may matter there.

(31) No aliens saw doves.

Mental models According to mental models theory of human reasoning, the first step in drawing a linguistic inference involves constructing a mental representation (= a mental model) of the premise based on its meaning and world knowledge (Johnson-Laird & Bara, 1984, a.o.). In order to decide whether a sentence (28) entails (31), reasoners might attempt to construct a mental model in which (28) is true and (31) is false. Finding such a model would result in the judgment that (28) does not entail (31), and not finding it in the judgment that (28) entails (31).

Such mental representations may include a finite set of aliens, a finite set of birds, and the seeing relations between the two sets. How may NPIs influence monotonicity inferences? According to (30), the mental representation of (29) may include more birds than the mental representation of (28). If each subset of birds has non-zero likelihood of being 'left out' of the mental model of (28), this means that inferences of the form 'No aliens saw x', for x a subset of birds, are more likely to follow from (29) than from (28), because a mental model in which some alien saw x is more likely to be compatible with (28) than with (29).

Interestingly, because the use of NPIs invites one to construct essentially a more complete and more accurate mental model, a general prediction of this approach is that NPIs should improve inferences. This means that in NM environments, their effect should be that they help participants perceive them as both less DE and less UE. How does this relate to our findings of the influence of NPIs on monotonicity inferences in NM environments? According to the results in Section 8.2, NPIs (i) reduce UE-ratings and (ii) increase DE-ratings in NM environments. The finding (i) is compatible with the prediction that NPIs should improve inferences (as NM environments do not support upward inferences), but the finding (ii) isn't (as NM environments do not support downward inferences either). More work is thus needed to make the conjunction of mental models approach to human reasoning with the scalar theories of NPI licensing a viable explanation for the influence of NPIs on monotonicity inferences in NM environments uncovered in the present work.

Mental logic According to mental logic theories (Beth & Piaget, 2013, Braine, 1978, Rips, 1994 Rips, 1983, Braine & O'Brien, 1998, a.o.), human linguistic reasoning is a product of application of formal rules of inference on the logical skeleton of a sentence. A downward inference from (28) to (31) thus involves two steps. In the first step, the reasoner would recover the logical form of the two sentences; in the second step, they would attempt to derive a proof from (28) to (31) using formal rules of inference.

Let us see this using the version of mental logic proposed by Sippel & Szymanik (2018), which extends that of Geurts (2003). According to this proposal, the logical form of a sentence such as (28) is in (32a). A formal reasoning rule MON↓ (see Sippel & Szymanik (2018) for details on this and other inference rules) can be applied to (32a) if it is considered to be true that {x | x saw doves} ⊆ {y | y saw birds}, which is represented as (32b). This derives the conclusion in (32c), which corresponds to (31).

(32) a. NO↓ ({x | x is an alien}, {y | y saw birds})

b. ALL ({x | x saw doves}, {y | y saw birds})
c. NO↓ ({x | x is an alien}, {y | y saw doves})

How may NPIs influence monotonicity inferences? According to (30), it is in principle possible for someone to believe that {x | x saw doves} ⊂ {y | y saw *any* birds} and at the same time believe that {x | x saw doves}⊄{y | y saw birds}. The MON↓ rule will thus be more likely to apply to (29) than to (28) to derive (30), resulting in NPIs boosting downward inferences.

How may this extend to NM environments? NPIs could boost downward inferences in NM environments by the same mechanism if one erroneously believes that NM environments are DE, and that thus rules of inferences such as MON↓ may apply to them. It is not obvious to see however why the presence of NPIs would reduce the amount of upward inferences in NM environments under this approach.

Probabilistic approach to human reasoning According to probabilistic approaches to human reasoning, probability calculus rather than standard (propositional and predicate) logic calculus should be taken as the competence theory of human reasoning with quantified sentences (Chater & Oaksford, 1999). Deductive reasoning with quantified sentences according to these approaches starts with assigning probabilistic semantics to quantified sentences. For instance, (28) is interpreted as in (33), with $P(X|Y)$ standing for conditional probability of X given Y.

(33)

$$P(x \in \{z \mid z \text{ saw birds}\} \mid x \in \{y \mid y \text{ is an alien}\}) = 0$$

As individuals who saw doves are a subset of individuals who saw birds, $P(x \in \{z \mid z \text{ saw birds}\} \mid x \in \{y \mid y \text{ saw doves}\}) = 1$. With this fact and (33) as premises, using laws of probability calculus, one can derive that $P(x \in \{z \mid z \text{ saw doves}\} \mid x \in \{y \mid y \text{ is an alien}\}) = 0$. In other words, one derives inferences from superset to subset (downward inferences) in the scope of *no*.

How may NPIs influence monotonicity inferences? According to the scalar theories of NPI licensing, (30) holds, which entails (34).

(34) $P(x \in \{z \mid z \text{ saw } \textit{any} \text{ birds}\} \mid x \in \{y \mid y \text{ saw doves}\}) \geq P(x \in \{z \mid z \text{ saw birds}\} \mid x \in \{y \mid y \text{ saw doves}\})$

(34) leaves room for the NPI *any* in the scope of *no* to boost downward inferences from (29) to (31) as compared to from (28) to (31) if someone believes that $P(x \in \{z \mid z \text{ saw birds}\} \mid x \in \{y \mid y \text{ saw doves}\}) < 1$, which would allow for the possibility that $P(x \in \{z \mid z \text{ saw } \textit{any} \text{ birds}\} \mid x \in \{y \mid y \text{ saw doves}\}) > P(x \in \{z \mid z \text{ saw birds}\} \mid x \in \{y \mid y \text{ saw doves}\})$. We remain agnostic as to what factors may lead to this belief.

While this outlines why, at the competence level, NPIs might boost downward inferences in the scope of *no*, more work is needed to extend this to NPIs in NM environments, and to explain why the effect in NM environments is more robust than in DE environments. The initial step needed for this extension is providing a probabilistic semantics for sentences headed by NM quantifiers such as *exactly 12* and *only 12*. It is unclear to us, however, how this is to be done. We thus leave this as an open problem for probabilistic approaches to reasoning with quantified sentences.

10.3. PIs induce biases as a superficial and accidental frequency effect

A final possibility when it comes to which mechanism may be behind the influence of PIs on reasoning is that people statistically track logical properties of environments in which different lexical items appear, and that this information is a bias in reasoning tasks. For instance, as we encounter NPIs most often in DE environments and not in UE environments, this may bias us to perceive even NM environments as DE and not UE in the presence of an NPI.

While the first two possibilities for the mechanism behind the influence of PIs on monotonicity inferences discussed in Sections 10.1 and Section 10.2 are compatible only with the first family of approaches to NPI licensing (i.e. scalar approaches), this final possibility may be compatible with both families of approaches to NPI licensing discussed in Section 1.

If statistical tracking of distribution of lexical items is at play, one may explore whether there are other lexical items in language which are not PIs, but which nonetheless occur more frequently in environments with specific logical properties. If such items are found, the prediction of the statistical tracking hypothesis is that they too should have an influence on inferences that is comparable to that of PIs. This is a potentially interesting direction to explore in future work, as it may be revealing of an important source of bias in human reasoning.

11. Conclusion

In the four experiments reported in this paper, it was found that PIs affect reasoning judgments. These PI manipulations are subtle on the surface, and give rise to accordingly small and subtle effects on upward and downward monotonicity inferences. These effects were not found in previous investigations with simpler cases (plain DE and UE environments, see Szabolcsi et al., 2008), and likewise, we found that the effects are smaller or absent in those environments. However, in more complex cases where inferential judgments are more difficult, in particular in NM environments, PIs have been found to influence inferential monotonicity judgments across multiple experiments. These results thus reveal the influence on reasoning tasks of subtle and apparently minor choices of closed class words.

Interestingly, the current results also document the fact that monotonicity inferences may be rather difficult to assess in inferential judgments tasks (Geurts & van Der Slik, 2005, cf. also Appendix A). Whether this is simply a consequence of performing the experimental task in question is open for discussion, but surely it raises challenges for the question of how people derive and understand the truth-conditions of sentences for efficient communication, noting that as soon as one understands the truth-conditional meaning of a sentence, one should be able to see what is entailed by it. We note here that PIs can help filter out some misunderstandings, if they can be used by the speaker to signal DE-ness and not UE-ness.

The fact that PIs alter monotonicity inferences raises questions about how PIs interact with other aspects of language processing which 'run on' monotonicity. Many psycholinguistic phenomena which connect to monotonicity have been discussed in the literature: licensing of plural anaphoric reference (Kibble, 1997; Moxey & Sanford, 1986, 1993; Nouwen, 2003; Sanford & Moxey, 2004; Sanford, Moxey, & Paterson, 1994), processing difficulties (Clark, 1976, Deschamps, Agmon, Loewenstein, & Grodzinsky, 2015, Just & Carpenter, 1971), donkey anaphora interpretation (Kanazawa, 1994), among others. A question for future research is to investigate whether such phenomena are affected by the presence of PIs in relevant sentences.

Finally, we have discussed a number of options of which cognitive mechanisms may be behind the effect of PIs on inferences, and how such mechanisms connect to theories of PI licensing. One interesting possibility discussed is the connection between a language parser which assesses the monotonicity properties in relation to PI acceptability and monotonicity inference computation in reasoning tasks, another is the interaction between (scalar) semantics of PIs and reasoning processes, and yet another is statistical tracking of logical properties of environments in which different lexical items occur. This calls for developing richer formal models for how human reasoning works in the language modality. Such an ambitious project would be well-informed by fine-grained linguistic phenomena sensitive to inferential abilities, and its outcome would provide a clearer understanding of how linguistic abilities and human reasoning interact and possibly help each other.

Acknowledgements

We wish to thank Lewis Bott, Alexandre Cremers, Luka Crnič, Danny Fox, Philippe Schlenker, Benjamin Spector, Anna Szabolcsi, Lyn Tieu. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013) / ERC Grant Agreement n.313610 and n. STG 716230 CoSaQ, and was supported by ANR-17-EURE-0017 and by a travel grant from the ESF as part of the Euro-Xprag network.

Appendix A. Environment effect on monotonicity judgments

Across four experiments, we find robust effects of environment on monotonicity judgments. We report average UE-ratings and DE-ratings in different environments (disregarding the effect of PI) in Table A1, as they may be a useful reference point for future research.

Table A1
Mean DE-ratings and UE-ratings per environment (disregarding the effect of PI).

Monotonicity	Mean DE-rating (SD)	Mean UE-ratings (SD)
UE	30.4% (18.9%)	87.2% (15%)
DE	78.4% (18.2%)	33.8% (21.2%)
NM	28.5% (20.4%)	43.8% (27.8%)
DN	40.9% (23.7%)	55.6% (23.7%)

In the following analyses, we investigate the influence of environments on monotonicity inferences separately for two inferential directions: *superset/subset* (i.e. DE-ratings) or *subset/superset* (i.e. UE-ratings). To this end, we pool together the data from the four experiments, and subset it to items with the pair of environments of interest (e.g. *NM* vs. *DE*) with inference direction of interest (e.g. *superset/subset*). We fitted Bayesian linear mixed-effects regression models to participants' responses with Environment as predictor (e.g. *NM* vs. *DE*), with the maximal random-effect structures justified by the design. We used the default priors of the *brms* package (cf. Section 3.3).

Four sampling chains ran for 4000 iterations with a warm-up period of 2000 iterations for each model, resulting in 8000 samples for each parameter. See Section 3.3 for details on results reporting and interpretation.

When it comes to DE-ratings, there is (i) more than 99% posterior probability that NM environments are rated lower than UE environments ($\mathbb{E}(\mu) = -0.96$, CI $= [-1.56, -0.38]$, $P(\beta < 0) = .999$); (ii) 100% posterior probability that UE environments are rated lower than DN environments ($\mathbb{E}(\mu) = -5.58$, CI $= [-6.68, -4.47]$, $P(\beta < 0) = 1$); and (iii) 100% posterior probability that DN environments are rated lower than DE environments ($\mathbb{E}(\mu) = -19$, CI $= [-20.75, -17.31]$, $P(\beta < 0) = 1$).

When it comes to UE-ratings, there is (i) 100% posterior probability that DE environments are rated lower than NM environments ($\mathbb{E}(\mu) = -5.11$, CI $= [-6.21, -4.05]$, $P(\beta < 0) = 1$); (ii) 100% posterior probability that NM environments are rated lower than DN environments ($\mathbb{E}(\mu) = -6.30$, CI $= [-8.19, -4.40]$, $P(\beta < 0) = 1$); and (iii) 100% posterior probability that DN environments are rated lower than UE environments ($\mathbb{E}(\mu) = -15.28$, CI $= [-16.55, -13.99]$, $P(\beta < 0) = 1$).

There are a number of interesting observations to be made here.

First, NM environments are judged less UE than UE environments (even though surprisingly slightly less DE as well), while they are judged less DE and more UE than DE environments. They are thus largely perceived to be inbetween DE and UE environments when it comes to monotonicity judgments, which is consistent with previous findings (Chemla et al., 2011).

Second, DN environments are judged very differently from other UE environments (even though DN environments are logically UE). They are judged more DE and less UE than other UE environments, and more UE and less DE than DE environments. In other words, they too are perceived to be inbetween DE and other UE environments when it comes to monotonicity judgments. To our knowledge, monotonicity inferences have not been investigated systematically in such environments, and this is thus a novel observation.

Third, focusing on DN and NM environments, DN environments are perceived as both more UE and more DE than NM environments. This too is to our knowledge a novel observation.

Why DN and NM environments have the perceived monotonicity properties that they do is an interesting question. Potentially relevant factors include their syntactic or semantic complexity and their frequency. We hope that these data may thus invite future research into difficulties with monotonicity reasoning.

Appendix B. Environment instance-PI interaction per environment

The main goal of the study reported in this paper was to investigate the effect of PIs on monotonicity inferences. We observed such effects of PIs on inferences in NM and DN environments. Two instances of NM environments were tested: sentences headed by *Exactly 12* (cf. (20)), and sentences headed by *Only 12* (cf. (21)). Similarly, two instances of DN environments were tested: sentences with expressions *Every* and *not* (cf. (23)), and sentences with expressions *No* and *without* (cf. (24)). Is there a difference between the two instances of NM environments as per the effect of PI on monotonicity inferences, and similarly, is there such a difference between the two instances of DN environments? We report here supplementary analyses investigating this question.

We pool together the data from the four experiments, and subset it to items with either the PI of interest or no PI (e.g. *NPI* vs. *no PI*) in the premise in the environment of interest (e.g. *NM* environment). We fitted Bayesian linear mixed-effects regression models to participants directional ratings with the following predictors: PI (*present* vs. *absent*), Environment instance (corresponds to different environments of the same monotonicity; e.g. in the case of NM environments these are *Exactly 12* vs. *Only 12*), PI-Environment instance interaction term, and Inference direction (*superset/subset* vs. *subset/superset*), with the maximal random-effect structures justified by the design. We used the default priors of the *brms* package (cf. Section 3.3).

Four sampling chains ran for 10,000 iterations with a warm-up period of 5000 iterations for each model, resulting in 20,000 samples for each parameter. See Section 3.3 for details on results reporting and interpretation.

NPIs in NM: Exactly 12 vs. Only 12 When it comes to NPIs in NM environments, there is more than 95% posterior probability that the presence of PPIs in the premise decreases the directional ratings more in *Exactly 12* environment instance as opposed to *Only 12* environment instance ($\mathbb{E}(\mu) = -0.51$, CI $= [-1.10, 0.09]$, $P(\beta < 0) = .954$).

PPIs in NM: Exactly 12 vs. Only 12 When it comes to PPIs in NM environments, there is 42% posterior probability that the presence of NPIs in the premise increases the directional ratings more in *Exactly 12* environment instance as opposed to *Only 12* environment instance ($\mathbb{E}(\mu) = -0.06$, CI $= [-0.67, 0.53]$, $P(\beta > 0) = .419$).

NPIs in DN: Every-not vs. No-without When it comes to NPIs in DN environments, there is 13% posterior probability that the presence of NPIs in the premise decreases the directional ratings more in *Every-not* environment instance as opposed to *No-without* environment instance ($\mathbb{E}(\mu) = 0.71$, CI $= [-0.55, 1.96]$, $P(\beta < 0) = .131$).

PPIs in DN: Every-not vs. No-without When it comes to PPIs in DN environments, there is 90% posterior probability that the presence of PPIs in the premise increases the directional ratings more in *Every-not* environment instance as opposed to *No-without* environment instance ($\mathbb{E}(\mu) = 0.86$, CI $= [-0.56, 2.23]$, $P(\beta > 0) = .896$).

We find evidence that NPIs influence monotonicity inferences more (in that they decrease directional ratings more) in *Exactly 12* NM environment

instance than in *Only 12* NM environment instance. For completeness, we further investigate whether NPIs influence monotonicity inferences only in *Exactly 12* NM environment instance. We find that this is not the case: NPIs decrease directional ratings in both *Exactly 12* NM environment instance ($\mathbb{E}(\mu) = -2.15$, CI = [−2.97, −1.32], $P(\beta < 0) = 1$) and in *Only 12* NM environment instance ($\mathbb{E}(\mu) = -1.26$, CI = [−2.14, −0.38], $P(\beta < 0) = .996$). There is thus evidence that NPIs influence monotonicity inferences in both instances of NM environments, but that such influence is stronger in *Exactly 12* NM environment instance.

There is also some evidence, albeit weaker, that PPIs influence monotonicity inferences more (in that they increase directional ratings more) in *Every-not* DN environment instance than in *No-without* DN environment instance. For completeness, we further investigate whether PPIs influence monotonicity inferences only in *Every-not* DN environment instance. We find evidence that PPIs increase directional ratings in *Every-not* DN environment instance ($\mathbb{E}(\mu) = 2.92$, CI = [1.12,4.72], $P(\beta > 0) = .998$), as well as suggestive evidence, albeit weaker, that they have a similar effect in the *No-without* DN environment instance ($\mathbb{E}(\mu) = 1.19$, CI = [−0.46,2.84], $P(\beta > 0) = .924$). We thus tentatively conclude that PPIs influence monotonicity inferences in both instances of DN environments, but that such influence is stronger in *Every-not* DN environment instance.

We do not find evidence for differential influence of PPIs in different NM environment instances, nor of NPIs in different DN environment instances.

The question of what causes the contrasts in the strength of influence of NPIs on monotonicity inferences in different NM environment instances, and of PPIs in different DN environment instances, if such contrasts turn out to be robust, is left for future work.

References

Barker, C. (2018). Negative polarity as scope marking. *Linguistics and Philosophy, 41*(5), 483–510.

Beth, E. W., & Piaget, J. (2013). *Mathematical epistemology and psychology. 12*. Springer Science & Business Media.

Braine, M. D. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85*(1), 1.

Braine, M., & O'Brien, D. P. (1998). *Mental logic*. Psychology Press.

Burkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software, 80*(1).

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1).

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 38*(2), 191–258.

Chemla, E., Homer, V., & Rothschild, D. (2011). Modularity and intuitions in formal semantics: The case of polarity items. *Linguistics and Philosophy, 34*(6), 537–570.

Chierchia, G. (2006). Broaden your views: Implicatures of domain widening and the "logicality" of language. *Linguistic Inquiry, 37*(4), 535–590.

Chierchia, G. (2013). *Logic in grammar: Polarity, free choice, and intervention*. OUP Oxford.

Clark, Herbert H. (1976). *Semantics and comprehension*. The Hague: Mouton.

Crnič, L. (2014). Non-monotonicity in NPI licensing. *Natural Language Semantics, 22*(2), 169–217.

Denić, M. (2015). Minimizing scope ambiguity hypothesis. In *In Proceedings of ESSLLI 2015 Student Session* (pp. 54–65).

Denić, M., Chemla, E., & Tieu, L. (2018). Intervention effects in NPI licensing: a quantitative assessment of the scalar implicature explanation. *Glossa: a Journal of General Linguistics, 3*(1), 49.

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

Drenhaus, H., Graben, P. B., Saddy, D., & Frisch, S. (2006). Diagnosis and repair of negative polarity constructions in the light of symbolic resonance analysis. *Brain and Language, 96*(3), 255–268.

Drenhaus, H., Joanna, B., & Julianne, S. (2007). Some psycholinguistic comments on NPI licensing. In *, 11. Proceedings of Sinn und Bedeutung* (pp. 180–193).

Drenhaus, H., Saddy, D., & Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. In S. Kepser, & M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives, pages 145–165*. Berlin: Mouton de Gruyter.

Fauconnier, G. (1975). Polarity and the scale principle. *Chicago Linguistics Society, 11*, 188–199.

Gajewski, J. R. (2005). *Neg-raising: Polarity and presupposition*. PhD thesis. Massachusetts Institute of Technology.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition, 86*(3), 223–251.

Geurts, B., & van Der Slik, F. (2005). Monotonicity and processing load. *Journal of Semantics, 22*(1), 97–117.

Giannakidou, A. (1998). *Polarity sensitivity as (non) veridical dependency, volume 23*. John Benjamins Publishing.

Guerzoni, E. (2006). Intervention effects on NPIs and feature movement: Towards a unified account of intervention. *Natural Language Semantics, 14*(4), 359–398.

Herburger, E., & Mauck, S. (2007). A new look at Ladusaw's puzzle. In H. Zeijlstra, & J.-P. Soehn (Eds.), *78–84. Workshop on negation and polarity*.

Homer, V. (2020). Domains of polarity items. *Journal of Semantics, 38*(1), 1–48.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 16*(1), 1–61.

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

Kadmon, N., & Landman, F. (1993). Any *Linguistics and Philosophy, 16*(4), 353–422.

Kanazawa, M. (1994). Weak vs. strong readings of donkey sentences and monotonicity inference in a dynamic setting. *Linguistics and Philosophy, 17*(2), 109–158.

Kibble, R. (1997). Complement anaphora and monotonicity. In *Formal grammar* (pp. 125–136).

Krifka, M. (1995). The semantics and pragmatics of polarity items. *Linguistic Analysis, 25* (3–4), 209–257.

Ladusaw, W. (1979). *Polarity sensitivity as inherent scope relations*. PhD thesis. University of Texas Austin.

Lahiri, U. (1998). Focus and negative polarity in Hindi. *Natural Language Semantics, 6*, 57–123.

Makowski, D., Ben-Shachar, M. S., Chen, S. H., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology, 10*, 2767.

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

Moxey, L. M., & Sanford, A. J. (1993). *Communicating quantities: A psychological perspective*. Lawrence Erlbaum Associates, Inc.

Muller, H., & Phillips, C. (2018). *Negative polarity illusions, volume Oxford handbook of negation*. Oxford University Press.

Nicolae, A. C. (2017). Deriving the positive polarity behavior of plain disjunction. *Semantics and Pragmatics, 10*.

Nouwen, R. (2003). Complement anaphora and interpretation. *Journal of Semantics, 20* (1), 73–113.

Parker, D., & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition, 157*, 321–339.

Progovac, L. (2000). Negative and positive feature checking and the distribution of polarity items. *Negation in Slavic*, 88–114.

Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review, 90*(1), 38.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. MIT Press.

Rothschild, D. (2006). Non-monotonic NPI-licensing, definite descriptions, and grammaticalized implicatures. In *, 16. Proceedings of Semantics and Linguistics Theory* (pp. 228–240).

Saddy, D., Drenhaus, H., & Frisch, S. (2004). Processing polarity items: Contrastive licensing costs. *Brain and Language, 90*(1–3), 495–502.

Sanford, A. J., & Moxey, L. M. (2004). Exploring quantifiers: Pragmatics meets the psychology of comprehension. In *Experimental pragmatics* (pp. 116–137). UK: Palgrave Macmillan.

Sanford, A. J., Moxey, L. M., & Paterson, K. (1994). Psychological studies of quantifiers. *Journal of Semantics, 11*(3), 153–170.

Shao, J., & Neville, H. (1998). Analyzing semantic processing using event-related potentials. *Newsletter for the Center for Research in Language, 11*(5), 3–20.

Sippel, J., & Szymanik, J. (2018). *Monotonicity and the complexity of reasoning with quantifiers*. In Proceedings of the 40th Annual Meeting of the Cognitive Science Society NB.

Steinhauer, K., Drury, J. E., Portner, P., Walenski, M., & Ullman, M. T. (2010). Syntax, concepts, and logic in the temporal dynamics of language comprehension: Evidence from event-related potentials. *Neuropsychologia, 48*(6), 1525–1542.

Szabolcsi, A. (2004). Positive polarity–negative polarity. *Natural Language & Linguistic Theory, 22*(2), 409–452.

Szabolcsi, A., Bott, L., & McElree, B. (2008). The effect of negative polarity items on inference verification. *Journal of Semantics, 25*, 411–450.

Vasishth, S., Brussow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science, 32*(4), 685–712.

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language, 108*(1), 40–55.

Xiang, M., Grove, J., & Giannakidou, A. (2013). Dependency-dependent interference: NPI interference, agreement attraction, and global pragmatic inferences. *Frontiers in Psychology, 4*, 708.

Yanilmaz, A., & Drury, J. E. (2018). Prospective NPI licensing and intrusion in Turkish. *Language, Cognition and Neuroscience, 33*(1), 111–138.

Yurchenko, A., Den Ouden, D.-B., Hoeksema, J., Dragoy, O., Hoeks, J. C. J., & Stowe, L. A. (2013). Processing polarity: ERP evidence for differences between positive and negative polarity. *Neuropsychologia, 51*(1), 132–141.

Zwarts, F. (1995). Nonveridical contexts. *Linguistic Analysis, 25*, 286–312.