Glossa a journal of general linguistics

Denić, Milica, et al. 2018. Intervention effects in NPI licensing: A quantitative assessment of the scalar implicature explanation. *Glossa: a journal of general linguistics* 3(1): 49. 1–27, DOI: https://doi.org/10.5334/gjgl.388

RESEARCH

Intervention effects in NPI licensing: A quantitative assessment of the scalar implicature explanation

Milica Denić[1,2], Emmanuel Chemla[1] and Lyn Tieu[3,4]

[1] Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, 29 rue d'Ulm, 75005 Paris, FR

[2] Institut Jean-Nicod (ENS, EHESS, CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, 29 rue d'Ulm, 75005 Paris, FR

[3] Western Sydney University, Locked Bag 1797, Penrith NSW 2751, AU

[4] ARC Centre of Excellence in Cognition and its Disorders, Australian Hearing Hub, 16 University Avenue, Macquarie University NSW 2109, AU

Corresponding Author: Milica Denić (milica.denic@ens.fr)

This paper reports on five experiments investigating intervention effects in negative polarity item (NPI) licensing. Such intervention effects involve the unexpected ungrammaticality of sentences that contain an *intervener*, such as a universal quantifier, in between the NPI and its licensor. For example, the licensing of the NPI *any* in the sentence **Monkey didn't give every lion any chocolate* is disrupted by intervention. Interveners also happen to be items that trigger scalar implicatures in environments in which NPIs are licensed (Chierchia 2004; 2013). A natural hypothesis, initially proposed in Chierchia (2004), is that there is a link between the two phenomena. In this paper, we investigate whether intervention effects arise when scalar implicatures are derived.

Keywords: intervention effects; negative polarity items; scalar implicature; experimental semantics

1 Negative polarity items and intervention effects

Negative polarity items (NPIs) are expressions that are sensitive to the logical properties of the environment in which they occur. Examples of NPIs in English include *any*, *anybody*, *anywhere*, and *ever*. A generalization that successfully captures the distribution of NPIs is that these items are acceptable in downward-entailing (DE) environments (Fauconnier 1975; Ladusaw 1979), i.e. environments that license inferences from sets to subsets. For example, (1a) entails (1b), with *chocolate muffins* denoting a subset of *muffins*, and the NPI *any* is acceptable. Conversely, the environment in (2) is not DE ((2a) does not entail (2b)), and the NPI *any* is not licensed.

(1) a. Ana didn't bake any muffins today.
    b. Ana didn't bake any chocolate muffins today.

(2) a. Ana baked (*any) muffins today.
    b. Ana baked (*any) chocolate muffins today.

Expressions that create a DE context for the NPI, such as negation, are called NPI licensors. Just as with negation in (1), we can see that *without*, in (3), and the restrictor of the universal quantifier, in (4), are also NPI licensors.

(3) a. John came to the party without any muffins.
b. John came to the party without any chocolate muffins.

(4) a. Everyone who tried any muffins that Ana baked loved them.
b. Everyone who tried any chocolate muffins that Ana baked loved them.

Certain DE environments, on the other hand, resist this generalization and do not license NPIs. This happens systematically when certain elements, so-called *interveners*, occur in between the licensor and the NPI, giving rise to so-called intervention effects (Linebarger 1987; Krifka 1995; Chierchia 2004; Beck 2006; Guerzoni 2006). For example, the universally quantified noun phrase (NP) is an intervener, but the definite NP is not: consider the two pairs in (5) and (6), which differ only in the intervener status of the indirect object. (5a) and (6a) entail (5b) and (6b), respectively, yet adding the NPI in (5a,b) leads to degradation, while adding the NPI in (6a,b) does not.

(5) a. Monkey didn't give every rabbit (*any) juice.
b. Monkey didn't give every rabbit (*any) strawberry juice.

(6) a. Monkey didn't give the rabbits any juice.
b. Monkey didn't give the rabbits any strawberry juice.

Another intervener is the conjunction *and*: when conjunction is in the scope of negation, (7a) entails (7b), yet the NPI *any* is not licensed.

(7) a. Ana didn't bake both cookies and (*any) muffins.
b. Ana didn't bake both cookies and (*any) chocolate muffins.

Strikingly, interveners form a natural class: they are items that trigger so-called scalar implicatures in DE environments (Chierchia 2004). Roughly, scalar implicatures are optional inferences arising when (i) a sentence can be argued to have a minimally different *alternative*, obtained for instance by replacing some lexical item with a similar one; for instance, *some* and *all*, or *or* and *and* would count as similar, or *scale-mates* (see Horn 1972; Gazdar 1979, as well as Katzir 2007 and Fox & Katzir 2011 for recent discussions about the derivation of alternatives), and (ii) the sentence is consistent with the alternative being false (see Grice 1975; Sauerland 2004; van Rooij & Schulz 2004; Schulz & van Rooij 2006; Spector 2006; 2007; Chierchia et al. 2008; Franke 2011; Bergen et al. 2016 for refinements and discussions from a variety of perspectives). In these conditions, the negation of the alternative may be added to the meaning of the sentence, as a scalar implicature. This is best understood through an example. (8a) can be argued to have (8b) as an alternative, obtained by replacing the universal quantifier *every* with the existential quantifier *any*. Because the environment is DE, the alternative (8b) is logically stronger than the sentence (8a), and therefore the negation of the alternative (8c) is compatible with the sentence, hence it may be added as a scalar implicature of the sentence.

(8) a. Sentence: Monkey didn't give every rabbit juice.
b. Alternative: Monkey didn't give any rabbit juice.
c. Scalar implicature: Monkey gave some of the rabbits juice.

Given that interveners are items that trigger scalar implicatures in DE environments, a natural hypothesis is that there is a link between the two. To make things slightly more concrete, one could say that scalar implicatures add a non-DE component to the meaning of a sentence, and for precisely this reason they may disrupt the licensing of an NPI.[1] The possible connection between implicatures and NPI licensing has been pursued by Chierchia (2004; 2013) in detail (for proposals that do not relate intervention effects to implicatures, see for example Beck 2006 and Guerzoni 2006). The proposal that implicatures are the cause of intervention effects is appealing in terms of its explanatory power. In particular, scalar implicatures are not the only inferences reported to give rise to intervention effects; Homer (2008) argues that presuppositions also give rise to intervention effects for the same reason, namely that they can disrupt the downward-entailingness of the licensing environment.

In this paper, we will focus on the proposal that implicatures might be the cause of intervention effects in sentences like (5). The question is whether such intervention effects arise when scalar implicatures are derived. An immediate reason to doubt that this is the case is that implicatures tend to be volatile, while intervention effects are reported to create categorically bad sentences. To explain the purportedly categorical ungrammaticality of intervention sentences, Chierchia (2004) proposes that the strong meaning of a sentence, i.e. its meaning enriched with an implicature, is available automatically in parallel with its plain implicature-less meaning, and crucially, it is this *strong* meaning against which NPI licensing must be checked.[2]

We will start by assessing whether the initial objection to the implicature-based theory is empirically justified to begin with. We will develop appropriate methods to measure intervention effects and assess what potential volatility they themselves may exhibit (Experiment 1). We will observe that they do not lead to categorical ungrammaticality as usually assumed, which reduces the initial challenge to the implicature theory, as it actually connects the two volatile phenomena.

The results of Experiment 1 are thus a motivation to abstract away from the specific details of Chierchia's proposal regarding the exact licensing mechanism of NPIs and how it interacts with implicature derivation. Instead, we will consider the expectations that arise from a possibly more general family of theories according to which implicatures are the cause of intervention effects, hereafter referred to as the scalar implicature (SI) theory of intervention effects. In particular, we will propose to evaluate whether the variability of one can be traced to the variability of the other.

[1] More precisely, the relevant implicatures in typical intervention configurations may turn a DE environment into an overall non-monotonic environment, i.e. an environment that does not license inferences from sets to supersets or subsets. Example (i) involves a non-monotonic environment (albeit not by way of implicatures); it is known that NPIs in such environments are not perfectly acceptable (see Rothschild 2006 and Crnič 2014 for discussion, and Chemla et al. 2011 for quantitative data).

(i) a. Exactly two boys baked muffins. ⇏ Exactly two boys baked chocolate muffins.
b. Exactly two boys baked muffins. ⇏ Exactly two boys baked.

[2] In light of this issue of the automaticity of strong (implicature-enriched) meanings, an anonymous reviewer asks whether processing data on implicatures may be relevant. In fact, the question of whether implicatures are derived automatically or not is orthogonal to the question of whether variability in implicature derivation can explain variability in intervention effects. The question only becomes relevant if (i) implicatures are derived automatically *and* (ii) one stipulates that NPI licensing must be checked before any implicature cancellation can occur. If this were the case, however, intervention effects would be expected to be categorical, and not volatile. As we will see, the results of our Experiment 1 run counter to this expectation. Regarding the processing of implicatures, we will merely note here that the majority of previous experimental work suggests that interpreting a sentence without its implicatures is equivalent to not deriving the implicatures rather than cancelling them; for instance, Bott & Noveck (2004) and Bott et al. (2012) provide evidence that participants who give implicature-less responses are faster to respond than those who give implicature-based responses, and Cremers & Chemla (2014) report similar results for *indirect* scalar implicatures, which are the kind that concern us here (though see also Romoli and Schwarz 2015).

To do so, we will make use of a range of common paradigms to compare spontaneous implicature derivation and grammaticality judgments of intervention sentences within individual participants. To preview, a rather mixed empirical picture will emerge across the different experiments. Experiments 2 and 3, using a picture selection task and a covered picture task to measure implicature derivation, reveal no correlation between individual implicature derivation rates and sensitivity to intervention effects. Experiment 4, employing a training paradigm to train participants to derive or not to derive implicatures, likewise reveals no correlation between implicatures and intervention. Finally, Experiment 5, which tests participants *simultaneously* on implicatures and intervention by observing their repair strategies for intervention sentences, reveals evidence that people may react to intervention sentences by effectively shutting off implicatures. As we will discuss in Section 7, the mixed empirical landscape revealed by the full set of experiments raises new challenges for theories of intervention.

The data and R analysis scripts (R Core Team 2016) for the experiments are available online at http://semanticsarchive.net/Archive/2U2ODU3N/Denic-Chemla-Tieu-InterventionEffects.html.

2 Experiment 1: Grammaticality judgments of intervention effects

As was pointed out in Section 1, the SI theory of intervention effects is challenged by the fact that scalar implicatures can be suspended, while sentences with the NPI in an intervention configuration are reported to be categorically ungrammatical. The goal of Experiment 1, which used an acceptability judgment task, was two-fold: confirm experimentally the presence of intervention effects in NPI licensing, and assess whether intervention configurations indeed lead to categorical ungrammaticality, as is usually assumed.

2.1 Participants

54 participants (15 female) were recruited on Amazon Mechanical Turk, and were paid $1.80 for their participation. One participant was excluded from analysis for not being a native speaker of English.

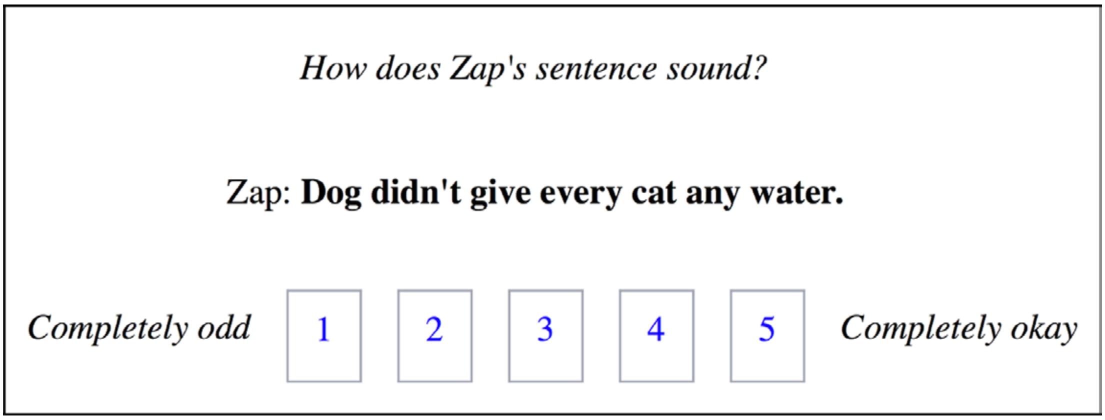
2.2 Procedure and materials

Participants were directed to a web-based acceptability judgment task, hosted on Alex Drummond's Ibex platform for psycholinguistic experiments.[3] Participants were told that they would read sentences (about animal characters) produced by Zap, an alien learning English, and that they were to judge how these sentences sounded on a scale from 1 to 5, with 1 being completely odd and 5 being completely okay. Participants registered their ratings by clicking on the appropriately numbered box (see Figure 1 for an example of the response buttons).

Participants first saw two practice trials, one involving a clearly well-formed sentence and one involving a clearly ill-formed sentence, accompanied by suggested ratings of 5 and 1, respectively. The purpose of these examples was to demonstrate to the participants that Zap was indeed capable of producing both acceptable and odd sentences. Participants then began the test phase of the experiment, the first two items of which were identical to the two practice trials. These were then followed by the 48 test trials schematically described in Table 1. We manipulated two factors, crossing NPI (present vs. absent) with the six item types listed in Table 1. The 48 trials were presented in randomized order.

As seen in Table 1, in addition to the target sentences, participants saw sentences involving DE and non-monotonic environments, as well as sentences involving different kinds

[3] The instructions for all experiments can be found in Appendix A; the full list of experimental items is provided in Appendix B.



How does Zap's sentence sound?

Zap: **Dog didn't give every cat any water.**

Figure 1: Experiment 1: An example of a test item with the NPI *any* in an intervention configuration.

Table 1: Summary of trial types.

Item type	Linguistic environment	Example sentence	Repetitions
Target	Negation, +intervener	Monkey didn't give every bear (any) pie.	16 [+NPI], 4 [–NPI]
Downward-entailing (DE)	Negation, –intervener	Monkey didn't give the rabbits (any) juice.	4 [+NPI], 4 [–NPI]
Non-monotonic (NM)	Scope of *exactly two*	Exactly two elephants ate (any) cake.	4 [+NPI], 1 [–NPI]
Upward-entailing (UE) – Simple	Positive sentence	Lion drank (any) coffee.	4 [+NPI], 1 [–NPI]
Upward-entailing (UE) – Complex	Nuclear scope of *every*, –negation in restrictor	Every rabbit who was hungry ate (any) chocolate.	4 [+NPI], 1 [–NPI]
Upward-entailing (UE) – Illusory	Nuclear scope of *every*, +negation in restrictor	Every bear who didn't have ice cream ate (any) pie.	4 [+NPI], 1 [–NPI]

of upward-entailing (UE) environments, which license inferences from sets to supersets. The UE sentences were not expected to license the NPI, and included simple UE sentences (UE-Simple), more complex UE sentences (UE-Complex), and UE sentences with a DE operator but in an irrelevant position (UE-Illusory). The particular number of repetitions of each of the non-target trial types was chosen in such a way as to balance them in terms of the overall number of degraded sentences (i.e. UE-simple, UE-complex, UE-illusory, NM[+NPI]) and non-degraded sentences (ie. DE[+NPI], all [–NPI] items).

The target sentences involved negated ditransitive constructions built on the following structural template: <Subject NP> didn't give every <Indirect Object NP> <Direct Object NP>. For the subject and indirect object NPs, we randomly chose animals from the following list: lion, rabbit, cat, dog, giraffe, bear, monkey, elephant. The direct object NPs were randomly chosen from the following list: juice, tea, milk, ice cream, pie, cheese, cake, honey. We opted to use mass nouns, in order to avoid any confound related to the plurality of objects.

2.3 Results

2.3.1 Exclusions

We excluded participants whose mean judgment for uncontroversially good cases (DE [+NPI] and all [–NPI] items) was *lower* than their mean judgment for uncontroversially bad cases (all UE [+NPI] items). This allowed us to ensure that participants were doing the task appropriately and understood the way the scale was supposed to be used. This criterion led to the exclusion of one participant, whose mean judgment on uncontroversially bad cases was 3.25, compared to a mean judgment of 2.69 for the uncontroversially good cases. As for the remaining participants, the mean judgment for the uncontroversially bad cases was 2.07, while the mean judgment for the uncontroversially good cases was 4.47.

2.3.2 Targets

Figure 2 presents the mean judgments across different environments from the remaining 52 participants (collapsing the three UE conditions); Figure 3 presents the same data in more detail for the critical sentences with an NPI. We are mostly interested in comparing the acceptability of the NPIs, in different environments. To do so, we computed a measure of this acceptability that factors out possible effects coming from the environment itself, independently of the NPIs. Specifically, for each participant and each environment, we calculated the log-ratio of the mean acceptability ratings for the [+NPI] targets over their [–NPI] counterparts, schematically: $\log\frac{\text{mean([+NPI])}}{\text{mean([-NPI])}}$. These log-ratios measure the degradation *due to the NPI*, for each participant in each environment; a comparison of log-ratios for two environments (below, through paired two-tailed t-tests) thus determines in which environment the introduction of an NPI yielded a higher degradation, i.e. which condition was judged worse.

First, we observe, as we expect, that NPIs are rated better in DE environments than in NM environments ($t(51) = 9.7$, $p < 0.001$), and that they are more acceptable in NM environments than in each of the UE environments (UE-simple: ($t(51) = 4.41$, $p < 0.001$, UE-complex: $t(51) = 3.61$, $p < 0.001$, UE-illusory: $t(51) = 2.45$, $p = 0.02$). This is as it should be, given the known intermediate status of NPIs in NM environments (Rothschild 2006; Chemla et al. 2011; Crnič 2014).

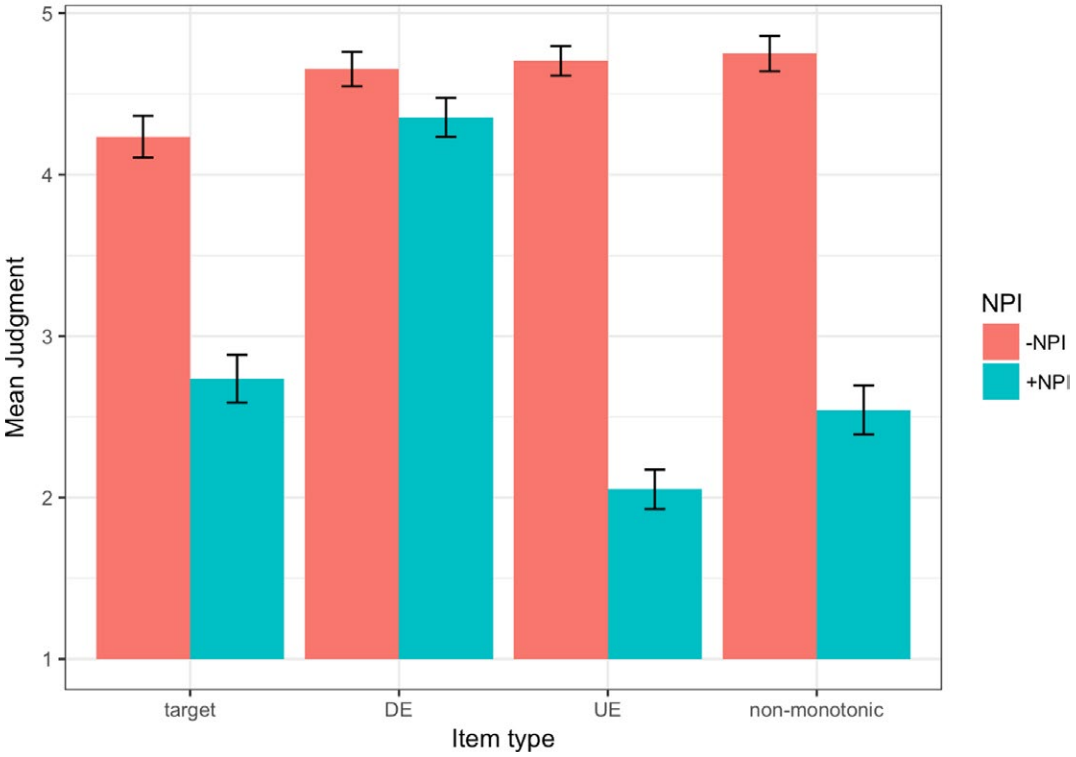


Figure 2: Experiment 1: Average response for each sentence type: target, downward-entailing, combined upward-entailing, and non-monotonic sentence types.

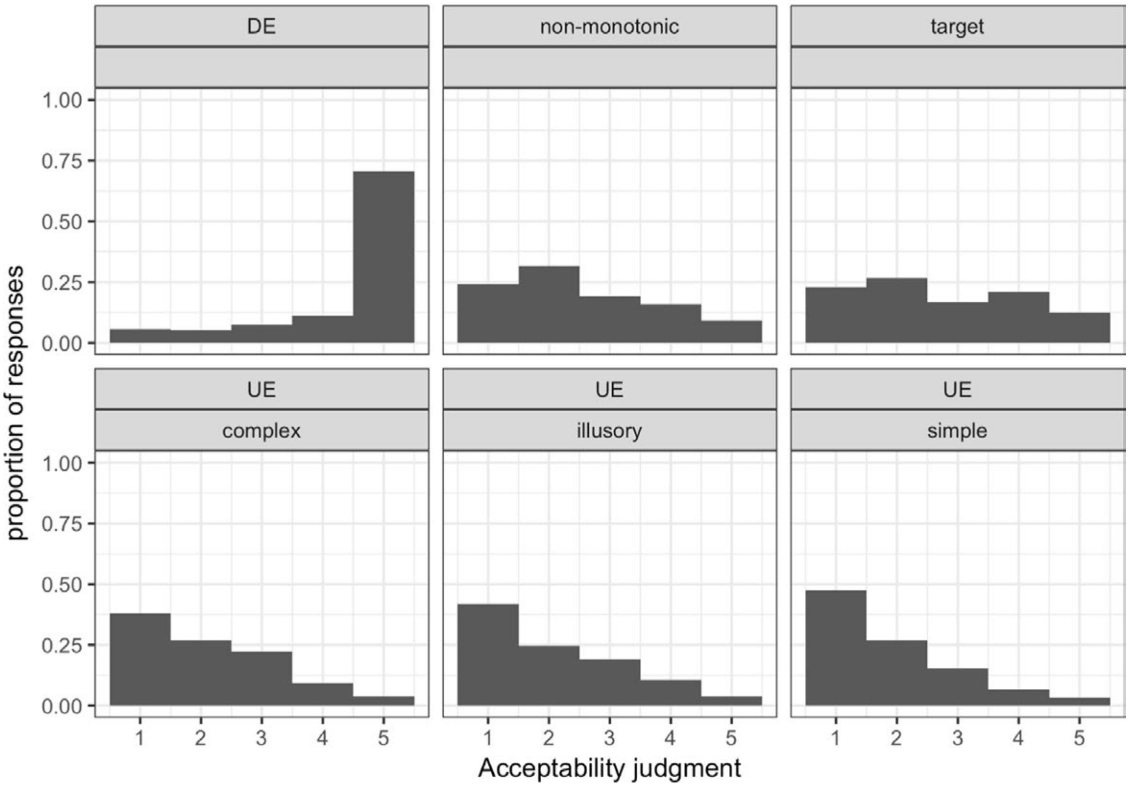


Figure 3: Experiment 1: Distribution of responses (1–5) for each sentence type in the [+NPI] condition.

Second, intervention effects were detected, in the sense that the critical target sentences were judged worse than the DE sentences ($t(51) = 6.8, p < 0.001$). Crucially, the degradation was not as strong as for plain violations in UE environments, with sentences in the target environments judged better than sentences in each of the UE [+NPI] conditions (UE-Simple ($t(51) = –7.9, p < 0.001$); UE-Complex ($t(51) = –6.5, p < 0.001$); UE-Illusory ($t(51) = –5.45, p < 0.001$)). Finally, the critical target sentences were also judged better than the NM sentences ($t(51) = –3.39, p = 0.001$).[4,5]

2.4 Discussion

Intervention effects were found to be NPI licensing violations, as the NPIs in the critical target sentences were judged as less acceptable than those in the DE sentences. But the violations that they yielded were weaker than other licensing violations, as they were still judged better than NPIs in the plain UE or NM sentences. This previously unnoticed intermediate status of intervention sentences in fact eliminates the initial objection to the implicature theory: that implicatures are volatile while intervention effects are not. This judgment pattern of intervention effects is thus interestingly compatible with the SI theory; intervention configurations may create NPI violations (say, because they create NM environments, see footnote 1), but this violation should only occur when the implicature

[4] An anonymous reviewer asks whether the target sentences might have been rated better than the NM sentences due to a sort of satiation effect, as participants saw more targets than they did NM items (cf. Table 1). In a post-hoc analysis, we observed that even when we restricted the analysis of the critical targets to the very first four occurrences, all of the significant differences remained.

[5] The reported p-values are reported unadjusted, so that the reader can check that the reported differences remain significant at various thresholds and with the application of Holm-Bonferroni corrections for multiple comparisons.

is derived. Testing the dependence of the violation on the presence of the implicature is the goal of the following experiments.

3 Experiment 2

In the first experiment it was shown that sentences containing NPIs in intervention configurations are judged better than sentences with NPIs in contexts that uncontroversially fail to license NPIs. This result goes well with the proposal that intervention effects are caused by scalar implicatures, but the experiment did not directly investigate the link between the two phenomena (because implicatures were not tested at all). Our next four experiments set out to investigate the presence of such a link.

We have seen that there is variation in the acceptability judgments of sentences involving intervention effects; that is, there is variation in the perceived *strength* of intervention effects (cf. target responses in Figure 3). Furthermore, previous experimental work has shown that people differ in how prone they are to derive scalar implicatures (Noveck & Posada 2003). The goal of Experiment 2 was to determine whether there is a relationship between the strength of intervention effects and the rate of derivation of scalar implicatures at an individual level. Such a relationship is expected under the SI theory, since not deriving the implicature should provide access to a grammatical parse of the sentence that would otherwise be considered ungrammatical due to the intervention effect.

3.1 Participants

54 participants (24 female) were recruited on Amazon Mechanical Turk and were paid $1.80 for their participation. Two participants were excluded from analysis for not being native speakers of English.

3.2 Procedure and materials

Experiment 2 involved two tasks: a Picture Selection Task (Roeper 2007) was used to estimate the participants' rates of implicature derivation, and an Acceptability Judgment Task was used to estimate the strength of the intervention effects. The order of the two tasks was counterbalanced across participants.

3.2.1 Acceptability Judgment Task

This task was almost identical to Experiment 1 in terms of the instructions and materials. There were two important differences, mostly imposed by the need to make room for a second task. First, there was no [–NPI] condition: all sentences in this task contained the word *any*. This prevented the possibility of evaluating the *contribution* of the NPI through the log-ratios as above, but to compensate for that we did not ask participants to report overall judgments about the sentences, but rather to report judgments about the contribution of the word *any*. Second, only the polar environments from Experiment 1 were tested: the target intervention environment, and one UE and one DE environment.

The Acceptability Judgment Task consisted of 19 items, which were preceded by three example items. The first item had an unambiguously unlicensed occurrence of *any*; in the immediate feedback, participants were told that most people would judge this item low on a scale from one to five. The second example item had a licensed *any* in the restrictor of a universal quantifier; the feedback to participants following this item was that most people would judge this item high on a scale from one to five.[6] The third example item was harder to judge; it contained *any* in a non-monotonic environment, which we know

[6] This example item did not include the same licensor (the restrictor of *every*) as the test items (negation), to avoid the possibility that participants would simply repeat (an abstract form of) the judgment they were given at the outset.

elicits variable judgments across speakers. Participants were told that some people would judge this item low and others high on the scale, and that they should simply follow their intuitions. The three examples were then presented again as the first three items of the experiment. The remaining 16 items consisted of eight target items and eight control items, presented in a pseudo-randomized order for each participant.

Target items were similar to the corresponding items from Experiment 1. An example target item is repeated in (9).

(9) Monkey didn't give every rabbit any juice.

There were four control items with a licensed *any* in the scope of negation, as in (10a), hereafter referred to as "good controls". The other four control items contained an unlicensed NPI *any*, as in (10b), hereafter referred to as "bad controls".

(10) a. Rabbit didn't drink any tea.
b. Lion drank any coffee.

3.2.2 Picture Selection Task

In the Picture Selection Task of Experiment 2, participants saw two pictures on each trial with a very short introduction sentence at the top of the screen that provided a setting for the story, and read a sentence that they were told had been produced by a puppet named Raffie.[7] Participants were instructed to evaluate Raffie's sentence with respect to the pictures on the screen. They were told that Raffie's sentence would sometimes be applicable to only one of the two pictures (Picture 1 on the left or Picture 2 on the right), and sometimes to both of the pictures. For each test sentence, participants were asked to decide from three response options: Picture 1, Picture 2, Both Pictures. An example item can be seen in Figure 4.

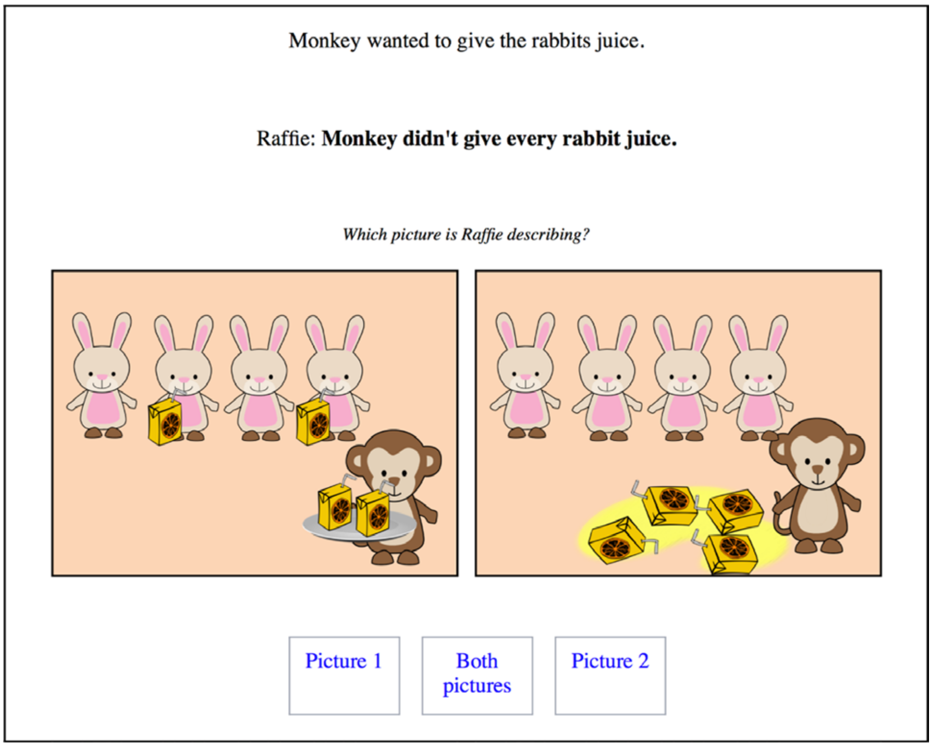
The Picture Selection Task consisted of 31 items. Participants were first presented with three examples. Raffie's sentence in the first example clearly matched Picture 1. Raffie's sentence in the second example clearly matched Picture 2. Raffie's sentence in the third example was harder to judge, and participants were told that some people would say it matched only one of the pictures, while others would say it matched both pictures, and they should simply follow their intuitions while doing the task. The three examples were then presented again as the first three items of the experiment. The remaining 28 items consisted of eight target items and 20 controls, presented in randomized order.

The target items in the Picture Selection Task were exactly parallel to those in the Acceptability Judgment Task, the only difference being that there was no NPI in the target sentences in the Picture Selection Task. An example target item from the Picture Selection Task is provided in Figure 4. In this example, the target sentence corresponds to (11a). Its logically stronger scalar alternative is in (11b). The scalar implicature of (11a) is thus the negation of (11b), i.e. the resulting interpretation of (11a) enriched with its scalar implicature is (11c).

(11) a. Sentence: Monkey didn't give every rabbit juice.
b. Alternative: Monkey didn't give any rabbits juice.
c. Overall meaning with SI: Monkey gave some but not every rabbit juice.

In Picture 1, Monkey gave two of the four rabbits juice, and in Picture 2 Monkey dropped all of the juiceboxes and none of the rabbits got juice. Now, if the participant derived the

[7] The materials for the study were made to be child-friendly, to allow for the possibility of a parallel child language acquisition study.



Monkey wanted to give the rabbits juice.

Raffie: **Monkey didn't give every rabbit juice.**

Which picture is Raffie describing?

Figure 4: Experiment 2: An example of a Picture Selection Task trial.

implicature, they would understand the sentence as (11c) (by the negation of the logically stronger alternative *Monkey didn't give any rabbits juice*). They would thus opt for the response "Picture 1". We will refer to this type of response as a *some*-response (as some of the animals in the picture received juice). On the other hand, if the participant didn't derive the implicature, they would interpret the sentence literally, true both in the situation in which Monkey gave only some of the rabbits juice (Picture 1) and in the situation in which he gave none of them juice (Picture 2); in this case, the participant would opt for the response "Both pictures".[8] Based on their response on target items, we can thus evaluate whether the participant interpreted the sentence with or without the implicature.

In addition to the targets, participants also received 20 control sentences. Four had the same construction as the target items, but their meaning was compatible with only one of the pictures. An example of this type of control is provided in (12a), which was paired with a picture in which two out of four elephants got cheese, and a picture in which four out of four elephants got cheese. Participants also saw four positive ditransitive control sentences, as in (12b), four negative ditransitive sentences with the universal quantifier in direct object position, as in (12c), and four negative ditransitive sentences with a definite noun phrase in indirect object position and a mass noun in direct object position, as in (12d). For half of the controls, the correct response was "Picture 1", and for the other half the target response was "Picture 2". Finally, there were four sentences like (12e), which were compatible with both of the pictures they were presented with.

[8] As an anonymous reviewer points out, participants who selected the response "Both pictures" might have been aware of both readings of the sentence: one with and one without the scalar implicature. For our purposes, however, it is not crucial to distinguish between participants who selected "Both pictures" because they only accessed the implicature-less reading and those participants who selected "Both pictures" because they had access to both the implicature- and implicature-less readings. What matters for us is that all of these participants in effect did have access to the implicature-less reading; under the implicature theory of intervention effects, all of these participants effectively had access to a grammatical parse of sentences containing an intervention configuration.

(12) a. Dog didn't give every elephant cheese.
b. Elephant gave every dog cake.
c. Giraffe didn't give Lion every strawberry.
d. Rabbit didn't give Cat tea.
e. Giraffe got every flower.

3.3 Results

3.3.1 Exclusions

Target sentences could be associated with two possible interpretations, one with the implicature and one without the implicature. There were therefore only two possible responses that participants could give: if the participant derived the implicature, they were expected to give a *some*-response (e.g., "Picture 1" in Figure 1); if the participant didn't derive the implicature, they were expected to select "Both pictures". Under no legitimate interpretation was a participant expected to select the image in which no animal gets the object in question (e.g., "Picture 2" in Figure 4); we will hereafter refer to these illegitimate responses as *none*-responses. *None*-responses would only be possible if the universal quantifier *every NP* could scope above negation, but this reading is normally unattested. Only one participant opted exclusively for such answers on our target items, so we decided to exclude them from the analysis. The remaining participants gave very few *none*-responses (6 out of 392), and these responses were excluded from the analysis as well.

Participants also had to correctly answer at least 75% of the controls in the Picture Selection Task to be included in the data analysis. This led to the exclusion of two additional participants.

For the Acceptability Judgment Task, participants' individual responses had to be such that, schematically: mean (good controls) ≥ mean (target) ≥ mean (bad controls). This requirement led to the exclusion of nine more participants, which left us with a total of 40 participants for analysis.

The 40 participants retained for the analysis responded with an average of 94.3% accuracy on control items in the Picture Selection Task. As for the controls in the Acceptability Judgment Task, the mean judgment for the good controls was 4.94 ($SD = 0.18$), and the mean judgment for the bad controls was 1.44 ($SD = 0.6$).[9]

3.3.2 SI derivation

On the Picture Selection Task targets, the proportion of responses consistent with an implicature (SI+ responses) was 0.56, with a standard deviation of 0.48.

3.3.3 Intervention measure

In the Acceptability Judgment Task, the critical target sentences received an average rating of 3.76 ($SD = 1.02$), which differed significantly from ratings for the good controls ($t(39) = –7.33, p < .001$) and the bad controls ($t(39) = 12.8, p < .001$).

3.3.4 Correlation

The main goal of Experiment 2 was to assess whether there was a correlation between implicature derivation and intervention effects.[10] First, we normalized each participant's responses on target items in the Acceptability Judgment Task by scaling each target

[9] Means and standard deviations (here and elsewhere) are calculated on within-participant means.

[10] We present all the results collapsed independently of which block occurred first: comparisons of mixed effects linear regression models (including random by-participant intercepts) with and without order of the two tasks as a fixed effect showed no effect of task order on intervention effects (χ^2 (1) $< 0.001, p = .98$). Similarly, the comparison of mixed effects logit models with and without the order of the two tasks as a fixed effect showed no effect of the order of the task on implicature derivation (χ^2 (1) $= 0.004, p = .94$).

response within the particular participant's extreme judgments of the good and bad controls, as in: $\frac{\text{target - mean}_p\text{ (bad)}}{\text{mean}_p\text{ (good) - mean}_p\text{ (bad)}}$, where mean$_p$ represents the mean of responses in a particular condition for a given participant *p*. This normalization corrects for different uses of the response scale by matching extreme values across participants. In Figure 5 below, each participant is represented at the height corresponding to the mean of these normalized responses for that given participant $\frac{\text{mean}_p\text{ (target) - mean}_p\text{ (bad)}}{\text{mean}_p\text{ (good) - mean}_p\text{ (bad)}}$. This is a value between 0 (intervention effects are as bad as bad controls) and 1 (intervention effects are as good as good controls), since we excluded participants who did not satisfy mean$_p$ (good) $\geq$ mean$_p$ (target) $\geq$ mean$_p$ (bad). To assess the correlation between the strength of intervention effects so represented and implicature derivation, we computed an implicature index for each participant: the proportion of SI+ responses to targets in the Picture Selection Task by that given participant. This implicature index is represented on the x-axis of Figure 5.

The first thing to note about this graph is that there is a wide range of perceived strength of intervention effects (participants are distributed all along the range of the y-axis), which replicates the finding from Experiment 1 that there is a lot of variation in the perception of intervention effects. The implicature index also shows variability, but in the form of bimodality (participants fall either to the far left or to the far right of the graph).

Turning to the correlation, mixed effects linear regression models were fitted to the normalized responses to the Acceptability Judgment Task targets, with mean response to the implicature targets as a fixed effect, and random by-participant intercepts. Comparisons with the models containing only the random by-participant intercepts revealed no significant effect of implicature response on the intervention judgments (χ^2 (1) = 0.35, p = .55). Experiment 2

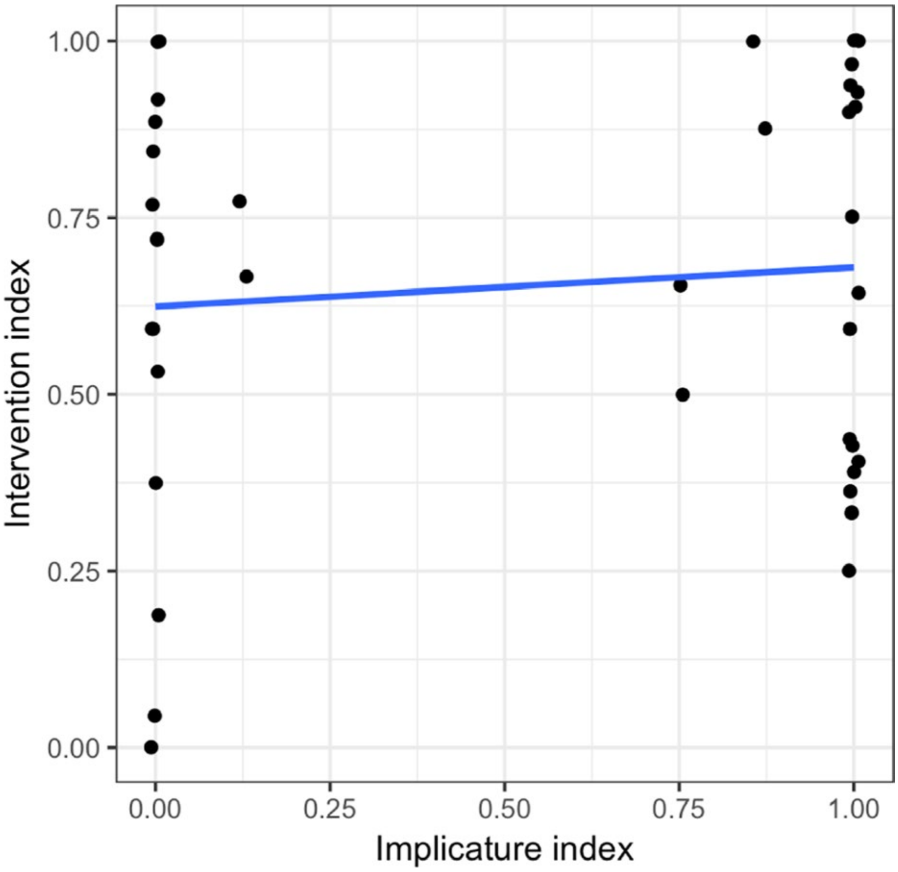


Figure 5: Results from Experiment 2: Individuals as a function of responses to implicature and intervention targets.

thus reveals no correlation between an individual tendency to derive implicatures and individual perception of intervention effects.[11]

3.4 Discussion

In this experiment, we measured participants' propensity to derive SIs and their sensitivity to intervention effects. While we observed variability in both, there was no observed correlation between the two indices.

4 Experiment 3

Experiment 3 had the same goal and general design as Experiment 2, except that it used a different task to estimate participants' propensity to derive scalar implicatures. In brief, we will see that Experiment 3 replicates the results of Experiment 2, revealing an absence of correlation between intervention effects and implicature derivation.

Participants in Experiment 2 were very consistent in the Picture Selection Task with respect to whether their responses were consistent with the implicature or consistent with an interpretation without the implicature. Given that intervention effects do not show such a robust bimodality, it is not surprising that SIs and intervention effects are not well correlated. However, the bimodality of the SI results may have been an artifact of the particular task we used to test for implicatures; thus in Experiment 3 we opted for a similar design but an alternative method for testing for SI derivation: the Covered Picture Task (see Huang et al. 2013).

4.1 Participants

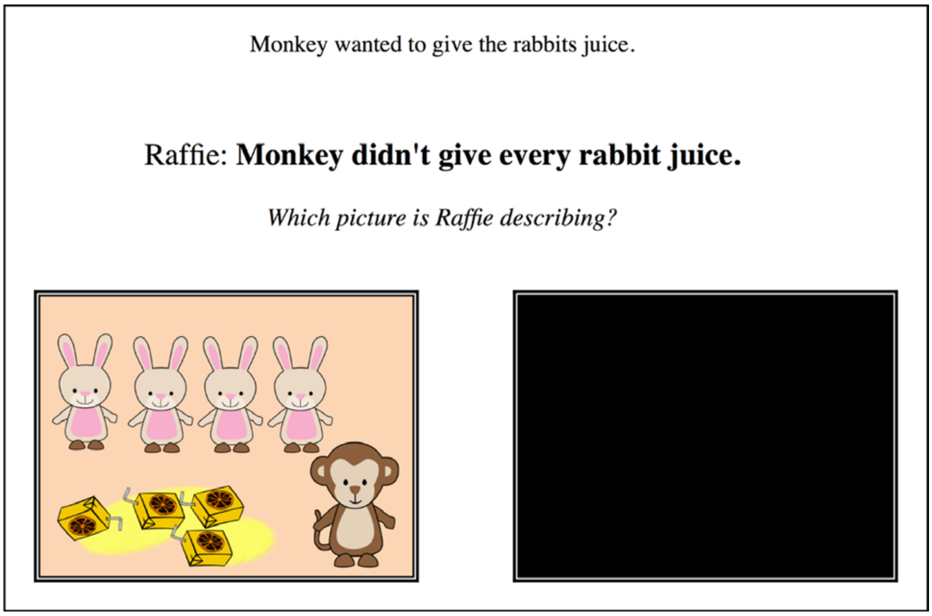
55 participants (27 female) were recruited through Amazon Mechanical Turk and were paid $1.80 for their participation. All participants reported English as their native language.

4.2 Procedure and materials

The procedure was the same as in Experiment 2, comprising two tasks. In addition to the Acceptability Judgment Task, participants completed a Covered Picture Task. In this task, they saw two pictures on the screen, one visible and the other covered, and a short introduction sentence at the top of the screen to introduce the story, followed by a target sentence produced by a puppet called Raffie. The participants were instructed to evaluate the visible picture with respect to Raffie's sentence. They were told that Raffie's sentence could apply to only one of the pictures — either the visible or the covered one. The participants had to click on the visible picture if they thought that Raffie's sentence was describing this picture, and on the covered picture if they thought that Raffie's sentence was describing a different picture from the visible one. An example of a target item is provided in Figure 6.

The visible picture that appeared with the target items was one that was incompatible with the derivation of the implicature. In the visible picture in Figure 6, we can see that none of the rabbits got juice. The participant would thus opt for the visible picture if they had access to the implicature-less reading, and for the covered picture if they expected, based on the target sentence, that at least some of the rabbits got juice, which is the

[11] One could investigate other predictions, more specific to the scalar implicature account. The standard view holds that the derivation of scalar implicatures requires extra processing time (see Bott and Noveck 2004 and much subsequent work; see Cremers & Chemla 2014 for similar results on the relevant *indirect* scalar implicatures, but also Romoli and Schwarz 2015 for an opposing view). If this is correct, then one would expect that low judgments of intervention effect sentences would require extra time in the Acceptability Judgment Task. Although this is borne out (χ^2 (1) = 3.67, p = .05), such an effect is not stronger than a similar effect under which lower judgments of ungrammatical sentences in general take more time; the interaction between the two types of effects is not statistically significant (χ^2 (1) = 0.99, p = .32).



Monkey wanted to give the rabbits juice.

Raffie: **Monkey didn't give every rabbit juice.**

Which picture is Raffie describing?

Figure 6: Experiment 3: An example of a Covered Picture Task trial.

reading of the target sentence enriched with an implicature, as in (11c). As in the Picture Selection Task of Experiment 2, we can evaluate based on responses to the target items whether the participant has interpreted the sentence with or without the implicature.

The Covered Picture Task contained 29 items. Participants were first presented with three examples. The first one involved a correct description of the visible picture; the immediate feedback to participants stated that in that case most participants would select the visible picture. The second example involved an incorrect description of the visible picture; here the feedback to participants was that most participants would select the covered picture. The third example was harder to judge; for this item, the participants were told that some people would select the visible picture and others the covered picture, and that they should follow their intuitions while doing the task. These three examples were then presented again as the first three items of the experiment. The remaining 26 items consisted of eight target items and 18 controls, presented in randomized order.

Among the 18 control items, 10 were compatible with the visible picture, and eight were not. They were very similar to the controls in the Picture Selection Task of Experiment 2 (but adapted for the Covered Picture Task).

4.3 *Results*

4.3.1 Exclusions

The exclusion criteria were the same as in Experiment 2, leading to the exclusion of 9 participants from analysis. The 46 participants retained for the analysis responded with an average of 92.1% accuracy on control items in the Covered Picture Task. In the Acceptability Judgment Task, the mean judgment for good controls was 4.84 ($SD = 0.42$), and the mean judgment for bad controls was 1.62 ($SD = 0.72$).

4.3.2 SI derivation

In the Covered Picture Task, the proportion of SI+ responses to the targets was 0.22, with a standard deviation of 0.31.[12]

[12] An anonymous reviewer points out that the estimate of implicature derivation in a covered picture paradigm may be noisy, as it incorporates biases across individuals, such as a bias against selecting the covered picture. To the extent that such a bias is not correlated with an individual participant's implicature derivation, with sufficient power it should not influence the main effect we are interested in.

4.3.3 Intervention effects

In the Acceptability Judgment Task, the critical targets received an average rating of 3.73 ($SD = 1.04$), which differed significantly from good controls ($t(45) = -8.39$, $p < .001$) and bad controls ($t(45) = 13.76$, $p < .001$).

4.3.4 Correlation

As in Experiment 2, we calculated each participant's intervention index, represented on the y-axis of Figure 7, and the proportion of implicature derivation, represented on the x-axis.[13] The closer the value on the x-axis to 1, the more implicatures the participant derived in the course of the Covered Picture Task; the closer the value on the y-axis to 1, the greater the judgments provided by the participant in response to the intervention targets in the Acceptability Judgment Task.

A mixed effects linear regression model was fitted to the normalized[14] responses to the Acceptability Judgment Task targets, with mean response to the implicature targets as a fixed effect and random by-participant intercepts. A comparison with the model containing only the random by-participant intercepts revealed no significant effect of covered picture implicature responses on the intervention judgments (χ^2 (1) $= 1.18$, $p = .28$).

4.4 Discussion

The results of Experiment 3 replicate those of Experiment 2: whether one uses a Picture Selection Task or a Covered Picture Task to estimate participants' rates of implicature derivation, we find that this estimate does not predict participants' acceptability judgments of intervention sentences.

5 Experiment 4: Training study

Experiments 2 and 3 revealed no correlation between participants' spontaneous rate of implicature derivation and their spontaneous grammaticality judgments of intervention sentences. In Experiment 4, we drew on previous experimental work revealing that people can be trained to derive implicatures or not to derive implicatures (see, for example, Noveck & Posada 2003; Bott & Noveck 2004; Chemla et al. 2017). We investigated whether training participants either to derive or not to derive implicatures would influence the strength of intervention effects, comparing the performance of a group that received training consistent with implicature derivation, with that of a group that received training inconsistent with implicature derivation.

5.1 Participants

54 participants (25 female) were recruited through Amazon Mechanical Turk and were paid $1.80 for their participation. All participants reported English as their native language.

5.2 Procedure and materials

The experiment was divided into two parts. The first task was designed to train half of the participants to derive SIs and the other half to not derive SIs. In the second part, participants received an Acceptability Judgment Task (designed to measure intervention effects), as in Experiments 2 and 3.

In the implicature training task, participants completed a Truth Value Judgment Task (TVJT) (Crain & Thornton 1998) (see Figure 8) in which they received immediate feedback

[13] Similar statistical methods as in Experiment 2 showed no effect of task order on either judgments of intervention effects (χ^2 (1) $= 0.86$, $p = .35$) or implicature derivation (χ^2 (1) $= 0.46$, $p = .5$). We therefore present all the results collapsed.

[14] Target responses were normalized in the same way as in Experiment 2.

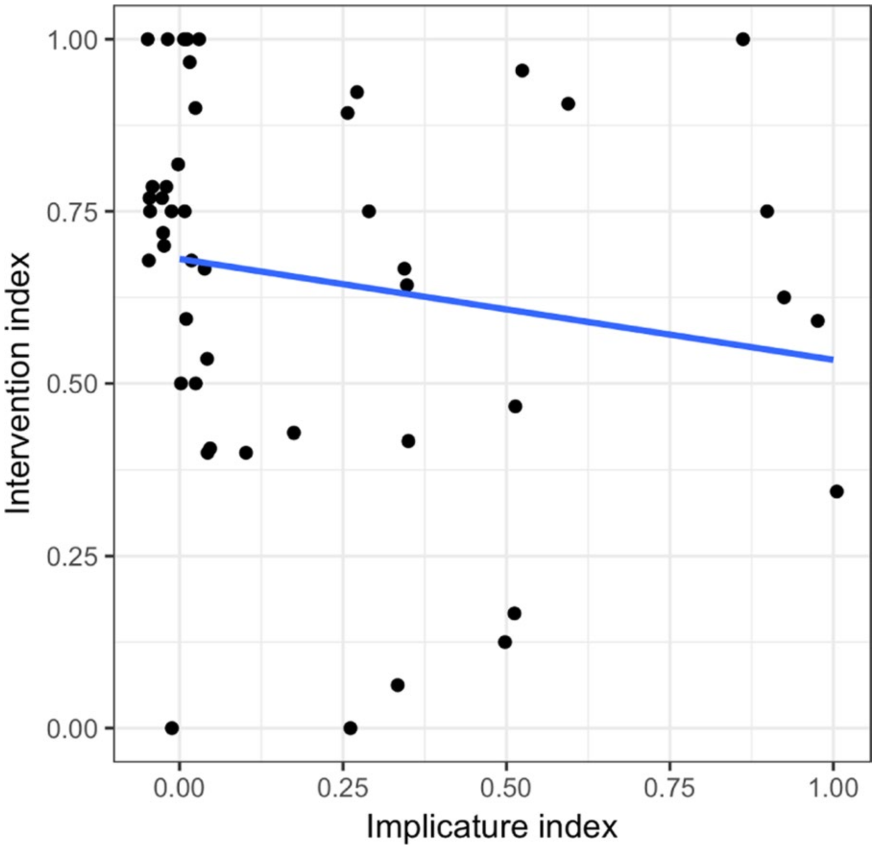
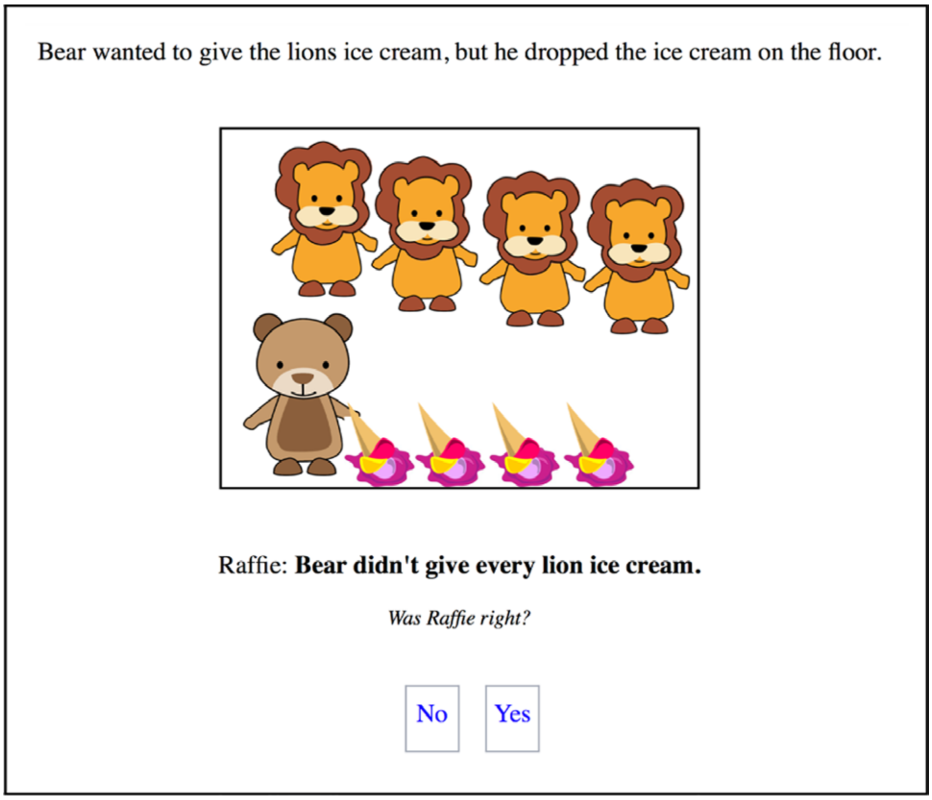


Figure 7: Results from Experiment 3: Individuals as a function of responses to implicature and intervention targets.



Bear wanted to give the lions ice cream, but he dropped the ice cream on the floor.

Raffie: **Bear didn't give every lion ice cream.**

Was Raffie right?

Figure 8: Experiment 4: An example of an implicature training TVJT trial.

after each trial. Participants saw a series of pictures with a short sentence at the top of the screen that introduced the story, followed by a target sentence produced by Raffie. Participants were instructed to evaluate the picture with respect to Raffie's sentence: they had to decide whether the descriptions were right or wrong. The response options were binary: the participants could either choose a "yes" or a "no" response.

If the participant derived the implicature, they would consider Raffie's sentence false in a situation in which none of the lions got ice cream (Figure 8), and would therefore select the "no" response. On the other hand, if the participant didn't derive the implicature, they would consider Raffie's sentence to be an acceptable description of the picture, and would select the "yes" response. What distinguished the group that was trained to derive implicatures (SI+ training group) from the group that was trained not to derive implicatures (SI- training group) was the feedback after the participant had provided a response on each of the target items. For example, for the experimental item in Figure 8, participants in the SI- training group received the feedback in (13), and participants in the SI+ training group received the feedback in (14).

(13) Feedback for SI- training group
a. "Yes" response: *Correct!*
b. "No" response: *Raffie was actually right — she said Bear didn't give every lion ice cream, and he didn't!*

(14) Feedback for SI+ training group
a. "Yes" response: *Raffie was actually wrong — she said Bear didn't give every lion ice cream, but none of the lions got ice cream!*
b. "No" response: *Correct!*

The task consisted of 28 items. Participants first saw two example items, the first a clearly true target and the second a clearly false target. For both example items feedback was provided about which response was correct. The example items were then presented again as the first two items of the implicature training TVJT. The remaining 26 items consisted of eight target items and 18 control items, presented in randomized order.

The target sentences on the implicature training TVJT were the same as those from Experiments 2 and 3. The pictures that accompanied these sentences were incompatible with the derivation of implicatures. The control items in this experiment were similar to those in Experiments 2 and 3. Of the 18 control items, 10 items had a clear "yes" target, and eight items had a clear "no" target. The main difference between these controls and those of the previous experiments was that the control trials in this experiment contained feedback, so as to be parallel with the target trials.

5.3 Results

5.3.1 Exclusions

The exclusion criteria were the same as in Experiments 2 and 3, leading to the exclusion of eight participants from analysis. The remaining 46 participants responded with an average of 95% accuracy on control items in the implicature training TVJT. In the Acceptability Judgment Task, the average judgment for good controls was 4.87 (*SD* = 0.34), and the average judgment for bad controls was 1.85 (*SD* = 0.99).

5.3.2 Training results

In the implicature training TVJT, the proportion of SI+ responses from the group of participants that received the SI+ training was 0.58 (*SD* = 0.25), while the group that received the SI- training gave no SI+ responses. This suggests that (i) the implicature-less response is salient from the start but (ii) the training is effective. To see this in more detail, Figure 9 displays the proportion of SI+ responses to the critical cases, arranged by the order in which they were presented. This graph allows us to visualize the effect of the SI+ training: all of the participants started off without implicatures, but they became more and more likely to derive the implicature with each subsequent SI+ training target.[15]

5.3.3 Intervention effects

In the Acceptability Judgment Task, participants who received the SI+ training gave the targets an average rating of 3.91 (*SD* = 0.84), which differed significantly from good controls (*M* = 4.93, *SD* = 0.21, *t*(23) = –6.33, *p* < .001), and from bad controls (*M* = 1.79, *SD* = 0.9, *t*(23) = 9.42, *p* < .001). Participants who received the SI- training gave an average rating of 3.71 (*SD* = 0.98), which differed significantly from good controls (*M* = 4.81, *SD* = 0.4, *t*(21) = –5.78, *p* < .001), and from bad controls (*M* = 1.92, *SD* = 1.13, *t*(21) = 6.91, *p* < .001).

A mixed effects linear regression model was fitted to the normalized[16] responses to the Acceptability Judgment Task targets, with the training received as a fixed effect, and random by-participant intercepts. A comparison with the model containing only the random by-participant intercepts revealed no significant effect of training received in the implicature training TVJT on subsequent acceptability judgments on the intervention items in the Acceptability Judgment Task (χ^2 (1) = 0.78, *p* = .38).

5.4 Discussion

The training administered in Experiment 4 succeeded in leading participants to either derive implicatures or to not derive implicatures. However, this training effect did not subsequently affect their acceptability judgments of the intervention targets. While the

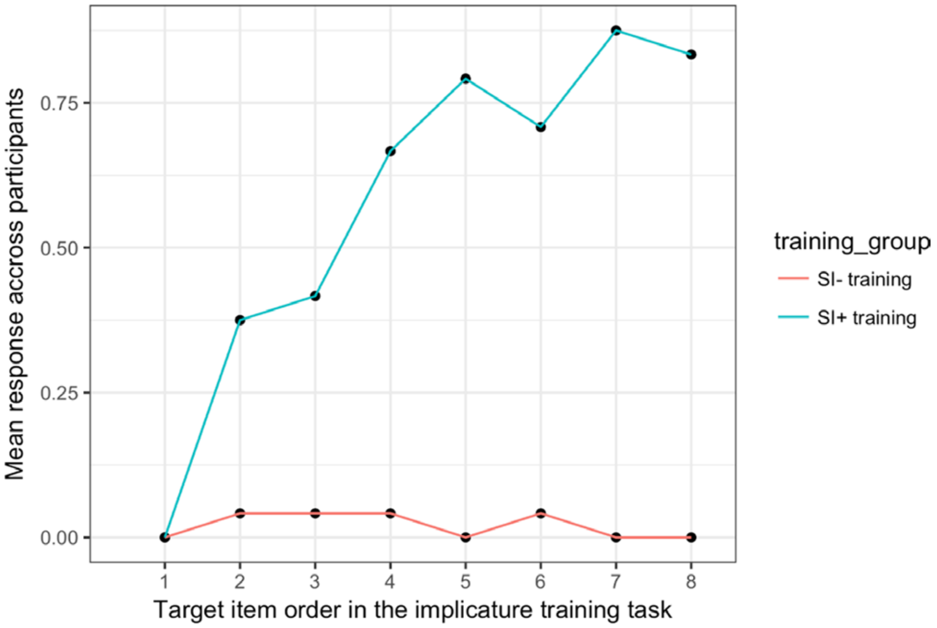


Figure 9: Experiment 4: Proportion of SI+ responses to implicature targets for each training group, by order of target appearance.

[15] An anonymous reviewer asks whether a similar increase in implicatures was observed in Experiments 2 and 3, which did not include a training component. We observed no such increase in implicatures across subsequent trials in Experiments 2 and 3, suggesting further that the SI+ training administered in Experiment 4 was indeed effective at encouraging participants to derive implicatures.

[16] Target responses were normalized in the same way as in Experiments 2 and 3.

SI theory might have led us to expect that the SI+ group would judge the intervention targets worse than the SI- group, this prediction was not borne out.

6 Experiment 5: Repair strategy study

Experiments 2–4 revealed no relationship between the rate of implicature derivation and the strength of intervention effects. This is surprising under the hypothesis that intervention effects are caused by the presence of scalar implicatures. The failure to observe this correlation could come from the fact that we measured SI derivation and strength of intervention at different points in time, while participants were carrying out different tasks. Hence, the SI derivation rates we measured may not be telling us much about whether participants derived implicatures *when they were providing judgments about the intervention effects*.

In Experiment 5, we thus attempted to measure the joint effects of implicature derivation and acceptability at a single point in time. The idea was the following. It is possible to assign a meaning to some sentences despite the fact that they are ungrammatical: we often and easily apply systematic "repair strategies" to non-grammatical sentences. The goal of Experiment 5 was to detect whether the suppression of the implicature is an available repair strategy for sentences containing intervention configurations. If the source of the ungrammaticality of the NPI in intervention configurations is the scalar implicature, a natural repair strategy would be to ignore the implicature and thereby allow the NPI to be licensed.

6.1 Participants

52 participants (24 female) were recruited through Amazon Mechanical Turk and were paid $1.80 for their participation. All participants reported English as their native language.

6.2 Procedure and materials

The experiment consisted of a single task, which was equivalent to the Picture Selection Task from Experiment 2. Participants were randomly assigned to one of two groups, which differed only in the target sentences they saw. The "regular" group saw the standard 8 [–NPI] target sentences from Experiment 2 (Appendix B.2, (6)). These sentences did not contain an intervention configuration, and therefore there was no need for any repair strategies. The "intervention" group saw almost the same 8 target sentences, except that these contained the NPI *any* in an intervention configuration. If participants in the intervention group detected an intervention effect, they were expected to *repair* the sentence in order to carry out the task. For example, the intervention group might see a target like that in Figure 10, while the regular group would see a target like Figure 4 from Experiment 2 (which is identical except that it does not include the NPI *any*).

In addition to the critical targets, both groups received the same 20 control items from the Picture Selection Task of Experiment 2 (Appendix B.2, (7)).

The intervention group (but not the regular group) was also asked to fill out a small questionnaire at the end of the experiment. In this questionnaire, the participants were asked to motivate their response choice on the target items: this questionnaire was there to make sure that participants had actually assigned a meaning to the target sentences, and had not simply responded at random.

6.3 Repair strategies

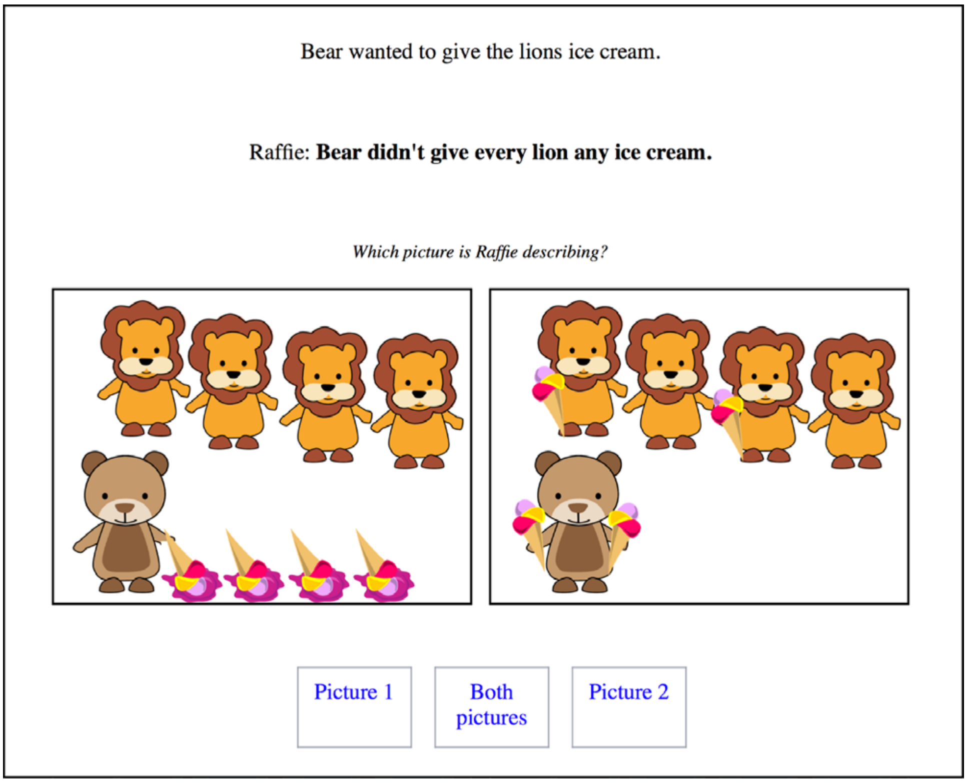
Before discussing the results, let us consider some possible repair strategies for a sentence like (15).

(15) *Bear didn't give every lion any ice cream.

One possibility would be to simply drop the NPI, and interpret the sentence as in (16a). Another possibility would be to reinterpret the universal quantifier as a definite description, as in (16b). Yet a third possibility would be to assign wide scope to the universally quantified noun phrase, above negation, as in (16c). The question of interest is whether there is a fourth possibility: to interpret the sentence as it is, *without the implicature*, as in (16d).

(16) a. Bear didn't give every lion ice cream.
b. Bear didn't give the lions any ice cream.
c. Every lion is such that Bear didn't give him ice cream.
d. Bear didn't give every (and possibly he didn't give any) lion ice cream.

What would be the expected responses on the Picture Selection Task under each of the above repair strategies? Table 2 summarizes possible repair strategies for (15) and their predictions, which we will now spell out. We will continue to refer to selections of pictures in which no animals got the object in question (e.g., a Picture 1 response in Figure 10)



Bear wanted to give the lions ice cream.

Raffie: **Bear didn't give every lion any ice cream.**

Which picture is Raffie describing?

Figure 10: Experiment 5: An example of a target item administered to the intervention group.

Table 2: Summary of repair strategies and the predicted responses.

Repair strategy	Post-repair interpretation	Prediction
Drop the NPI	(16a)	Same result pattern as in Experiment 2
Replace the universal quantifier with a definite description	(16b)	Increase of *none*-responses
Assign wide scope to the universal quantifier	(16c)	Increase of *none*-responses
Do not derive an implicature	(16d)	Increase of "Both pictures" responses

as *none*-responses,[17] and to selections of pictures in which some animals got the object in question as *some*-responses (e.g., a Picture 2 response in Figure 10).

Participants who adopt the strategy that leads to the interpretation in (16a), i.e. dropping the NPI, might respond in one of two ways, depending on whether they compute an indirect scalar implicature from (16a). If they compute the implicature *Bear gave some of the lions ice cream*, they are expected to give a *some*-response. If they do not compute an implicature from (16a), they are expected to select "Both pictures", since the implicatureless reading of (16a) is compatible with both pictures. Notice that if dropping the NPI is the strategy adopted by the intervention group, we may expect to see parallel performance in the two groups, since the reanalyzed (16a) is identical to the target sentences that the regular group is presented with.

Participants who employ the strategy of reinterpreting the universal quantifier as a definite description, as in (16b), or of giving the universal wide scope, as in (16c), would be expected to give a *none*-response. If the intervention group makes use of either of these two strategies, we expect to see a greater proportion of *none*-responses in the intervention group than in the regular group.

Finally, participants who employ the strategy of implicature suppression as in (16d) are expected to select "Both pictures". Since none of the above repair strategies lead to an increase in "Both pictures" responses, an increase of such responses in the intervention group could be interpreted as the propensity to employ this strategy (to be discussed further below).

6.4 *Results*

6.4.1 Exclusions

As in the earlier experiments, participants had to correctly answer at least 75% of the controls in order to be included in the data analysis. One participant failed to do so and was excluded from the analysis. The mean accuracy on controls for the remaining 51 participants was 94%.

6.4.2 Targets

Figure 11 displays the counts of the different response types for the implicature targets from the intervention group and the regular group. Recall that a "Both pictures" response corresponds to a non-implicature interpretation of the target sentence. To determine whether group (intervention vs. regular) had an effect on the type of response to the implicature targets, we recoded the responses in binary terms (as "Both pictures" and "Other"). We then fitted a mixed effects logistic regression model on these responses, with group as a fixed effect and random by-participant intercepts. A comparison of this model with a reduced model without the group fixed effect revealed a significant effect of group on response type (χ^2 (1) = 6.05, p = .01), with more "Both pictures" (non-implicature) responses in the intervention group than in the regular group.

6.5 *Discussion*

One interpretation of these results is that they are consistent with the SI theory: if scalar implicatures cause intervention effects, it is natural to expect that when faced with an NPI in an intervention configuration, participants may choose to repair the sentence by cancelling the derivation of the scalar implicature.

However, there are two alternative explanations for this effect. First, note that there is evidence that at least some of our participants repaired the target sentences by replacing

[17] Note that these responses should now be perfectly legitimate for the "intervention" group in the experiment.

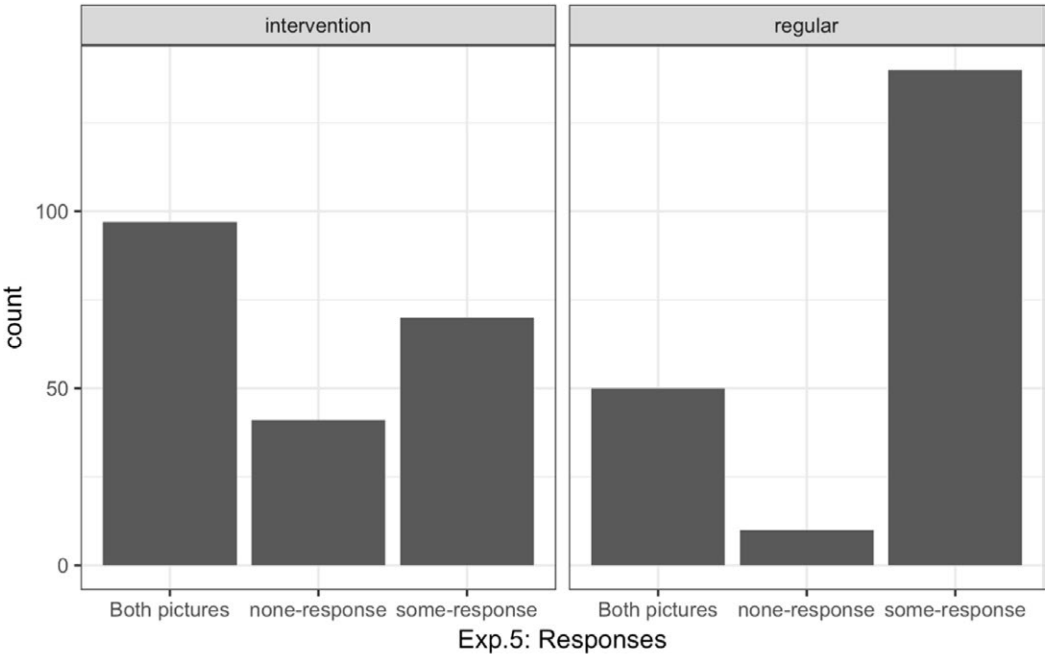


Figure 11: Experiment 5: Distribution of responses to implicature targets from the intervention and the regular group.

the universal quantifier with a definite description or by means of wide-scoping the universal quantifier: this is suggested by the increase in the amount of *none*-responses in the intervention group compared to the regular group. As these strategies seem to be available to some of our participants, the participants who selected "Both pictures" could have entertained not one but *two* possible repair strategies simultaneously, namely, one of the strategies leading to a *none*-response, in combination with the strategy of dropping the NPI. Such participants might simply not have known which of the two interpretations should be chosen. For example, such a participant might think that it is as probable that Raffie wanted to use an indefinite instead of the NPI, and derive the implicature (which would make the sentence compatible with Picture 2 of Figure 10), as it is that the universal quantifier should be assigned wide scope with respect to negation (which would make Raffie's sentence compatible with Picture 1 of Figure 10). The participants could therefore be unsure which of the two meanings Raffie intended to convey, which is why they might opt for "Both pictures" more often in the intervention group than in the regular group.

An argument against this explanation comes from the questionnaire that the intervention group was asked to complete at the end of the experiment. The explanations that participants gave for their picture selections can be found in Appendix B.5.2. These explanations confirm that for the most part, participants were not merely responding at random to the target items, but rather repaired the sentences to assign some meaning to them. The crucial question for us was whether participants who selected "Both pictures" might report confusion or uncertainty about the intended meaning of the sentence, and whether this uncertainty could have driven their picture selection. Upon close inspection, however, few responses were indicative of such confusion, with only one of the 10 "Both pictures" justifications referring to an ambiguity (i.e. *The sentence is a bit ambiguous and could either mean that the elephant didn't give any tea to the giraffes, or that he didn't give tea to all the giraffes*). The questionnaire responses therefore speak against an alternative explanation of the data according to which people were unsure about the intended meaning of the sentence due to the existence of more than one repair strategy.

While we do not find the multiple repair strategy explanation very strong in light of the questionnaire data, it is worth noting that we cannot at present completely rule it out either. Even when participants' explanations for "Both pictures" selections are compatible with the non-implicature reading (e.g., *In both pictures not every giraffe has tea, so both are correct*), we cannot definitively know whether they chose "Both pictures" because they did not derive the implicature, or because there were multiple repair strategies that would render the sentence true as soon as not all of the giraffes got tea.

A second alternative explanation is that participants in the repair group were under more cognitive load, which is why they derived fewer implicatures. Previous experimental work has shown that in a dual task setting people tend to prefer readings without scalar implicatures (De Neys & Schaeken 2007; Marty & Chemla 2013). Facing a broken sentence (with an intervention effect) and perhaps trying to interpret it could require additional cognitive load leading to fewer implicatures, not because implicature suppression is a possible repair strategy, but because participants are expending effort to repair the sentence.

Overall, one might have considered this experiment to hold the greatest chance of yielding an observable correlation between intervention effects and SIs, because it allowed both to be measured at a single point in time. Compared to the other experiments, this one yields the results most compatible with a positive conclusion in favor of a correlation, although there remain alternative explanations that require further investigation.

7 General discussion

The current studies provide the first quantitative measurement of the strength of intervention effects. Throughout all of the experiments, we observe that intervention effects are real, but not as strong as one might have expected.

A more specific goal of the present experiments was to use such measurements to evaluate the relationship between scalar implicature derivation and the presence of intervention effects in NPI licensing. The SI theory posits that intervention effects are caused by scalar implicatures. This raises a tension between the optionality of SIs and the categorical judgments reported in the theoretical literature on intervention effects. Experiment 1 suggests that one might be able to resolve this tension. In this experiment (as well as in subsequent experiments), intervention sentences were found to be judged better than different kinds of sentences with unlicensed NPIs, and there was great variability in how ungrammatical (if ungrammatical at all) people considered sentences with intervention configurations. This response pattern to intervention sentences goes well with the hypothesis that intervention effects are caused by scalar implicatures, as it could be that it is precisely the optionality of implicature derivation that leads to the variability in judgments of intervention sentences.

In Experiments 2 and 3, we thus investigated a more precise consequence of the implicature theory, whereby intervention effects should be contingent on the derivation of implicatures. But we observed no correlation between an individual's rate of implicature derivation and their sensitivity to intervention effects. In Experiment 4 we found that training people to derive or not to derive implicatures did not have an influence on their subsequent judgments of intervention effects. The results of Experiment 5, on the other hand, do suggest that when forced to assign a meaning to a sentence with an intervention configuration, participants behaved as though they had made it possible by blocking the derivation of scalar implicatures.

The present experiments thus paint a mixed empirical landscape. While we do not observe a relationship between judgments of intervention sentences and implicature derivation in Experiments 2–4, we do observe such a relationship in Experiment 5. Hence these results must be interpreted with care; if the implicature theory is on the right track, the absence of an effect in Experiments 2–4 must be explained. Note also that there is an important methodological difference between the experiments in which we observe null effects and the experiment in which we observe the expected effect, which may be quite relevant to explaining the findings.

In particular, Experiments 2–4 show that, if indeed intervention effects are caused by scalar implicatures, this will not manifest itself in a simple correlation between a task that tries to provide an individual's scalar implicature index, and a task that tries to provide an index of the individual's sensitivity to intervention effects. There could be multiple reasons for this; one possibility is that, as the two tasks are very different and implemented at different points in time, scalar implicature derivation in one task does not predict scalar implicature derivation later, in the other task.

Experiment 5, on the other hand, provided a *simultaneous* estimate of intervention effects and implicature derivation. This experiment is thus methodologically more powerful in terms of capturing the link between implicatures and intervention effects, if there indeed is one. These results must be interpreted with care, especially in light of the preceding absence of similar results in the previous experiments, but they are the best evidence of a link between implicatures and intervention effects.

Before closing, it is worth repeating that our study focused on testing a specific family of theories of intervention effects in NPI licensing, but there exist alternative accounts that do not appeal to scalar implicatures. For instance, Beck (2006) and Guerzoni (2006) relate the phenomenon to intervention effects in the domain of wh-words, where certain quantificational elements and operators like disjunction and conjunction can be seen to disrupt the licensing of wh-words as in (17) (examples adapted from Guerzoni 2006).

(17) a. *Which book did which student and Mary read?
b. *Which book did which student or Mary read?

While reviewing these alternative theories would take us beyond the scope of the present paper, we would point out that the experiments reported here still offer a very relevant challenge for alternative accounts, namely how to explain the variability in judgments of intervention sentences and the fact that they are judged to be better than sentences containing unlicensed NPIs. Building on the present study, one prediction of accounts like Beck (2006) and Guerzoni (2006) that could be investigated in future research is that we might observe a similar variability in the grammaticality judgments of intervention effects in wh-licensing. From where we stand, it's unclear how these alternative accounts would explain the variability observed in our study, as well as any variability one might observe in judgments of intervention effects in wh-licensing. An ideal theory of intervention effects should seek to explain such variability, and future work could seek evidence for the connection with wh-intervention in the same way as we have done here for the SI theory.

Of course, it could always be that intervention effects and the observed variability are due to non-grammatical factors. For instance, the intervention configurations may impose a particular burden on participants, making grammatical sentences resemble ungrammatical ones to some participants (see Gibson 1991; 1998 for related ideas about center embedding configurations or Sprouse et al. 2012 for related investigations on island effects).

Intervention could also be a real grammatical phenomenon, but one whose *variability* is due to individuals' limitations in independent processing abilities, e.g., working memory constraints. Even if such a scenario would take much of the task of explaining intervention out of the linguist's hands, it would be necessary to spell out and fully test such a proposal, in order to obtain a proper and complete understanding of intervention effects.

Additional Files

The additional files for this article can be found as follows:

- **Appendix A.** Consent form and instructions. DOI: https://doi.org/10.5334/gjgl.388.s1
- **Appendix B.** Stimuli. DOI: https://doi.org/10.5334/gjgl.388.s1

Acknowledgements

We would like to thank Maria Aloni, Amir Anvari, Gennaro Chierchia, Luka Crnič, Anna Czypionka, Jakub Dotlačil, Manuel Križ, Jeremy Kuhn, Mora Maldonado, Salvador Mascarenhas, Rick Nouwen, Philippe Schlenker, and Benjamin Spector for helpful discussion, critiques, and questions, and Dorothy Ahn for allowing us to use her artwork in our experiments. We are also grateful to three anonymous reviewers for their helpful comments, questions, and suggestions. The research leading to this work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.313610, by ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC, and by the Australian Research Council Centre of Excellence in Cognition and its Disorders (CE110001021).

Competing Interests

The authors have no competing interests to declare.

References

Beck, Sigrid. 2006. Intervention effects follow from focus interpretation*. *Natural Language Semantics* 14(1). 1–56. DOI: https://doi.org/10.1007/s11050-005-4532-y

Bergen, Leon, Roger Levy & Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* 9. DOI: https://doi.org/10.3765/sp.9.20

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

Bott, Lewis, Todd M. Bailey & Daniel Grodner. 2012. Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language* 66(1). 123–142. DOI: https://doi.org/10.1016/j.jml.2011.09.005

Chemla, Emmanuel, Chris Cummins & Raj Singh. 2017. Training and timing local scalar enrichments under global pragmatic pressures. *Journal of Semantics* 34(1). 107–126. DOI: https://doi.org/10.1093/jos/ffw006

Chemla, Emmanuel, Vincent Homer & Daniel Rothschild. 2011. Modularity and intuitions in formal semantics: The case of polarity items. *Linguistics and Philosophy* 34(6). 537–570. DOI: https://doi.org/10.1007/s10988-012-9106-0

Chierchia, Gennaro. 2013. *Logic in grammar: Polarity, free choice and intervention*. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199697977.001.0001

Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In Adriana Belleti (ed.), *Structures and beyond* 3. 39–103. Oxford University Press.

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2008. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In Klaus von Heusinger, Claudia Maienborn & Paul Portner (eds.), *Semantics: An international handbook of natural language meaning*, New York, NY: Mouton de Gruyter.

Crain, Stephen & Rosalind Thornton. 1998. *Investigations in universal grammar*. Cambridge, MA: MIT Press.

Cremers, Alexandre & Emmanuel Chemla. 2014. Direct and indirect scalar implicatures share the same processing signature. In Salvatore Pistoia Reda (ed.), *Pragmatics, semantics and the case of scalar implicatures* (Palgrave Studies in Pragmatics, Language and Cognition), 201–227. London: Palgrave Macmillan. DOI: https://doi.org/10.1057/9781137333285_8

Crnič, Luka. 2014. Non-monotonicity in NPI licensing. *Natural Language Semantics* 22(2). 169–217. DOI: https://doi.org/10.1007/s11050-014-9104-6

De Neys, Wim & Walter Schaeken. 2007. When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental psychology* 54(2). 128–133. DOI: https://doi.org/10.1027/1618-3169.54.2.128

Fauconnier, Gilles. 1975. Polarity and the scale principle. *Chicago Linguistics Society* 11. 107–126.

Fox, Danny & Roni Katzir. 2011. On the characterization of alternatives. *Natural Language Semantics* 19(1). 87–107. DOI: https://doi.org/10.1007/s11050-010-9065-3

Franke, Michael. 2011. Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics* 4. 1–82. DOI: https://doi.org/10.3765/sp.4.1

Gazdar, Gerald. 1979. *Pragmatics, implicature, presuposition and logical form*. Academic Press.

Gibson, Edward. 1991. *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Pittsburgh, PA: School of Computer Science, Carnegie Mellon University dissertation.

Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68(1). 1–76. DOI: https://doi.org/10.1016/S0010-0277(98)00034-1

Grice, H. Paul. 1975. Logic and conversation. *Syntax and Semantics,* 41–58.

Guerzoni, Elena. 2006. Intervention effects on NPIs and feature movement: Towards a unified account of intervention. *Natural Language Semantics* 14(4). 359–398. DOI: https://doi.org/10.1007/s11050-007-9008-9

Homer, Vincent. 2008. Disruption of NPI licensing: The case of presuppositions. In *Proceedings of SALT* 18. DOI: https://doi.org/10.3765/salt.v18i0.2483

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

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

Katzir, Roni. 2007. Structurally-defined alternatives. *Linguistics and Philosophy* 30(6). 669–690. DOI: https://doi.org/10.1007/s10988-008-9029-y

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

Ladusaw, William A. 1979. *Polarity sensitivity as inherent scope relations*. Austin, TX: University of Texas at Austin dissertation.

Linebarger, Marcia C. 1987. Negative polarity and grammatical representation. *Linguistics and Philosophy* 10(3). 325–387. DOI: https://doi.org/10.1007/BF00584131

Marty, Paul & Emmanuel Chemla. 2013. Scalar implicatures: Working memory and a comparison with 'only'. *Frontiers in Psychology* 4(403). DOI: https://doi.org/10.3389/fpsyg.2013.00403

Noveck, Ira A & Andres Posada. 2003. Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language* 85(2). 203–210. DOI: https://doi.org/10.1016/S0093-934X(03)00053-1

R Core Team. 2016. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.Rproject.org/.

Roeper, Thomas. 2007. *The prism of grammar: How child language illuminates humanism*. MIT Press.

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

Rothschild, Daniel. 2006. Non-monotonic NPI-licensing, definite descriptions, and grammaticalized implicatures. In *Proceedings of SALT* 16. 228–240. DOI: https://doi.org/10.3765/salt.v16i0.2944

Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27(3). 367–391. DOI: https://doi.org/10.1023/B:LING.0000023378.71748.db

Schulz, Katrin & Robert van Rooij. 2006. Pragmatic meaning and nonmonotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy* 29(2). 205. DOI: https://doi.org/10.1007/s10988-005-3760-4

Spector, Benjamin. 2006. *Aspects de la pragmatique des opérateurs logiques*. Paris, France: Université de Paris 7 dissertation.

Spector, Benjamin. 2007. Scalar implicatures: Exhaustivity and Gricean reasoning. In Maria Aloni, Alastair Butler & Paul Dekker (eds.), *Questions in dynamic semantics*, 225–249. Elsevier. DOI: https://doi.org/10.1163/9780080470993_011

Sprouse, Jon, Matt Wagers & Colin Phillips. 2012. A test of the relation between working memory capacity and syntactic island effects. *Language* 88(1). 82–123. DOI: https://doi.org/10.1353/lan.2012.0004

van Rooij, Robert & Katrin Schulz. 2004. Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information* 13(4). 491–519. DOI: https://doi.org/10.1007/s10849-004-2118-6

How to cite this article: Denić, Milica, Emmanuel Chemla and Lyn Tieu. 2018. Intervention effects in NPI licensing: A quantitative assessment of the scalar implicature explanation. *Glossa: a journal of general linguistics* 3(1): 49. 1–27, DOI: https://doi.org/10.5334/gjgl.388

Submitted: 26 March 2017 **Accepted:** 14 January 2018 **Published:** 20 April 2018

]u[*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press. **OPEN ACCESS**