Direct and indirect scalar implicatures share the same processing signature*

Alexandre Cremers† Emmanuel Chemla†

Abstract

Following the seminal work of Bott and Noveck (2004), investigations into the psycholinguistic properties of scalar implicatures (SIs) have mostly focused on *direct* SIs, e.g. when a sentence with 'some' is understood as negating a stronger alternative with 'all' ("Some x are y" implies that it is not the case that "All x are y"). Most previous studies found that SIs incur a processing cost. In this study, we investigate *indirect* SIs, i.e. implicatures which arise when a sentence with 'all' is understood as negating an alternative with 'some'. This typically happens in negative sentences ("Not all x are y" implies some x are y), for negation reverses entailment relations between sentences. We report on two truth-value judgement tasks designed to compare direct and indirect SIs. In Exp. 1, we found the traditional cost observed for direct SIs but not for indirect SIs. However, in Exp. 2 we show that once effects of negation are factored out, the two classes of SIs can be seen to share the same processing properties. Hence, there is a cost inherent to SIs and it generalizes across different subclasses of the phenomenon. This "signature" of SIs should now be compared with other kinds of inferences, either to understand these inferences and their relation to SIs (Chemla and Bott, 2011, 2012) or to better identify which subprocesses specifically involved in the derivation of an SI are responsible for this cost.

1 Linguistic and psychological approaches to scalar implicatures

1.1 Linguistic Background

1.1.1 What are scalar implicatures?

Scalar implicatures (henceforth SIs) are inferences which arise when a speaker utters a sentence like (1):

*We wish to thank Lewis Bott, Celena Carden-Ribault, Kyle Duarte, Shiri Lev-Ari, Roger Levy, Jenner Martin, Paul Marty, Philippe Schlenker, Greg Scontras, Benjamin Spector for helpful remarks and criticisms. We would like to thank Alex Drummond for the existence of Ibex and his availability to discuss related issues with us. 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 was supported by ANR-10-IDEX-0001-02 and ANR-10-LABX-0087. This version is a pre-publication draft.

†LSCP – ENS Paris

(1) Some of the students came to the party.

This utterance can be understood as (2a) or (2b):

(2) a. Some or all of the students came.
b. Some, but not all, of the students came.

(2a) is strictly weaker than (2b). (2b) adds the inference that "not all students came", which is called a scalar implicature. The "not all" scalar implicature in (2b) is optional, as shown by the fact that it can be felicitously cancelled in (3a). By contrast, if a similar "not all" inference was not a scalar implicature but a logical consequence of the literal meaning of the sentence, it may not be cancelable, as shown by the infelicity of (3b).

(3) a. Some of the students came to the party. In fact they all came.
b. # None of the students came to the party. In fact they all came.

From the work of Grice (1967), a derivation has been proposed for such inferences which goes in substance as follows:

1. If a speaker utters the sentence with 'some', the addressee may wonder why she did not use the sentence with 'all': "All of the students came to the party."
2. This alternative is in fact more informative (it entails the sentence with 'some')[1].
3. The addressee assumes that the speaker does not believe the stronger alternative to be true, otherwise she would have uttered it instead of the sentence with 'some'.
4. If the addressee further assumes that the speaker is *opinionated* about the alternative (i.e. the speaker knows whether it is true or false)[2], then she can conclude that the alternative is false: not all students came to the party.

This reasoning does not only apply to the competition between 'some' and 'all'. Whenever two terms systematically compete, we can define a *scale*. The example above involves the scale <some, all> but there are more: <or, and>, <might, must>,<warm, hot>, etc.

1.1.2 Indirect scalar implicatures

The theory presented above makes predictions for any sentence in which a term from a scale appears. We showed how sentences with 'some' can be compared with their alternative with 'all', but the reverse is also possible. Any two sentences which differ only in the use of 'some' or 'all' can be compared and several situations are conceivable: They can be logically equivalent, one can be strictly stronger than the other, and they can be independent. In the simplest cases (e.g., 4a), the sentence with 'all' is stronger, but it is not always the case. In (4b) for instance, the sentence with 'all' is not stronger than the sentence with 'some' (or 'any', which is the equivalent of 'some' in negative contexts).

[1] Actually, as soon as the alternative is not less informative an implicature arises. See Spector (2003); van Rooij and Schulz (2004) for formal accounts and Chemla and Spector (2011); Chemla et al. (2011) for experimental confirmation of this claim.

[2] This assumption is sometimes referred to as the "epistemic step" (Sauerland, 2004; van Rooij and Schulz, 2004; Spector, 2005).

(4) Alternatives and their entailment relations

a. John saw **all** of the students ⇒ John saw **some** of the students

b. John didn't see **all** of the students ⇏ John didn't see **any** of the students

("⇒" represents material implication: Whenever the left member holds, so does the right member)

In cases like (4b) the sentence with 'all' triggers an SI, which is the negation of the stronger alternative with 'some/any'.

(5) Evaluation of possible scalar implicatures from sentences with 'all'

a. John saw **all** of the students ⇝̸ John didn't see any of the students

b. John didn't see **all** of the students ⇝ John saw some of the students

("⇝" stands for "gives rise to the inference")

These inferences are derived following the same reasoning as for the 'not all' inference in (1). For instance, the inference in (5b) corresponds to the negation of the alternative in (4b), i.e. "It is not the case that John didn't see **any** of the students", which is equivalent to "John saw some of the students".

In simple affirmative cases like (4a), 'all' yields a stronger sentence than 'some', and for this reason it is sometimes called the strong element of the scale <some, all>. Environments which reverse this relation, such as (4b) are called *downward entailing environments*. As shown in (5), these environments trigger SIs when 'all', rather than 'some', occurs in them. Such SIs are sometimes called *indirect*, whereas implicatures triggered by the use of a 'weak' element in a scale are called *direct* SIs.

Our focus in the rest of this study will be on the comparison between direct SIs in simple, affirmative environments and indirect SIs in negative sentences.

1.2 Psychological background

If deriving a scalar implicature requires the computation of alternatives and the negation of some of these alternatives as described in the previous paragraph, then SIs should have a processing cost. One of the first series of experiments to test this prediction was presented in Bott and Noveck (2004) (see also Noveck and Posada, 2003). They proposed to measure the derivation cost of SIs using a sentence-verification task with sentences such as (6). This sentence is true under its weak reading (without an SI, as in (2a)) but false under its strong reading (with an SI, as in (2b)):

(6) Some elephants are mammals.

a. Weak reading: Some or all elephants are mammals. (True)

b. Strong reading: Some but not all elephants are mammals. (False)

Overall, the results show that when participants answer *False* (that is according to the strong reading with an SI) participants were significantly slower than when they answer *True* (that is according to the weak reading without an SI). Subsequently, many studies, using different procedures, confirmed that giving an answer associated with a strong reading to a sentence containing a scalar item is costly (Huang and Snedeker, 2009a,b; Degen and

Tanenhaus, 2011; Bott et al., 2012). This cost has been attributed to the derivation process of scalar implicatures.

Nevertheless, it is still unclear what the source of this cost is. Some cost may come from the additional complexity of the meaning contributed by the inference. However, recent results have controlled for complexity by using explicit paraphrases of the overall meaning (using phrases such as 'only some', as in Bott et al., 2012) and proved that there is a residual cost beyond that. Furthermore, increasing the working memory load with a double task reduces the rate at which participants derive SIs, but does not affect their responses to truth conditionally equivalent 'only some' sentences (Marty and Chemla, 2013).

1.3 Why study indirect SIs?

Most experimental studies of scalar implicatures have focussed on a specific instance of the phenomenon. An exception to this rule is found in the work of van Tiel et al. (2013), who compared the derivation rates of SIs between a variety of scales. They showed that there is great variability in the SI derivation rates across scales. Our goal is also to extend the empirical explored territory, and we will do so by looking at online processes, focussing on potential differences between direct and indirect scalar implicatures.

The appeal of the theories of SIs is that they explain a wide class of inferences with the same basic principles. The theories mostly relies on alternative generation (see Katzir, 2007's approach and van Tiel et al., 2013's investigations of different scales) and a routine to take alternatives into account. In particular, nothing specific needs to be said about indirect SIs once the theory is in place for direct SIs. We will thus use indirect SIs to test whether the conclusions of Bott and Noveck (2004) generalize to the routine and in particular whether the processing cost is a core property of this SI derivation mechanism or a mere accident associated to a subclass of scalar implicatures.

Studying indirect SIs also addresses some methodological issues. Chemla and Bott (2011, 2012) used a similar paradigm to study other phenomena (presuppositions, free-choice inferences) and argued that they did not share a common derivation mechanism with SIs because they displayed reversed processing signatures, despite theoretical claims that had been made to unify these phenomena (Kratzer and Shimoyama, 2002; Fox, 2007; Chemla, 2009; Romoli, 2013). Indirect SIs in contrast are uniformly derived like direct SIs in the theoretical literature, but they usually involve some form of negation. Negation may have dramatic effects on sentence-verification tasks (Clark and Chase, 1972; Carpenter and Just, 1975), so direct and indirect SIs provide a good example of phenomena which share a common derivation mechanism but display some superficial differences which may affect the processing signature. In order to validate the conclusions from the studies mentioned above it is crucial to test such examples: If the paradigm cannot disentangle superficial differences in the sentences from differences in the derivation of the inferences, we could not conclude that the phenomena studied in Chemla and Bott (2011, 2012) must be explained by different mechanisms. More generally, if several classes of SIs turn out to share the same processing pattern, despite superficial differences, then we may be more confident in using this pattern as a signature of SIs.

In the following, we report on 2 experiments that aimed at comparing the processing of

direct and indirect SIs.

2 Experiment 1: *à la Bott & Noveck*

2.1 Goal of this experiment

This experiment was designed to investigate the time-course of indirect scalar implicatures in a design similar to Bott and Noveck (2004). It was a sentence-verification task in which target sentences were true under the weak reading and false under the strong reading. Participants received no explicit training or instructions on how to treat these target sentences. By doing so, we were able to see how often participants would spontaneously derive SIs. We included both direct and indirect targets, so we were able to compare the derivation rates of both classes of SIs. Provided average derivation rates sufficiently close to 50%, we would also be able to compare the response times associated with weak and strong readings for each class of SIs.

2.2 Methods and Materials

2.2.1 Course of the experiment

The experiment consisted in a sentence-verification task: Participants read sentences and had to judge whether they were true or false.

Participants were recruited on Amazon's Mechanical Turk and redirected to an online experiment hosted on Alex Drummond's Ibex Farm. Response times were recorded locally and sent to the server at the end of the experiment, so that participants' internet connection would not interfere with the measure. Before starting the experiment, participants received instructions and read examples of true and false sentences. They also read one example of an ambivalent sentence with the scalar item "or". They were told that such sentences are intermediate and that opinion about their truth may vary. In such cases they would have to follow their intuition.

After reading the instructions, participants started the experiment. They were asked to use the keys **D** (*False*) and **K** (*True*) to answer.

The first five sentences were practice items and included examples presented in the instructions. They were designed to help participants familiarize with the task and were not included in the analyses. The remaining items were presented in random order.

At the end of the experimental phase, participants had to fill a questionnaire. This questionnaire included questions about their age, sex and native language, and the device used to answer (keyboard, mouse, touchscreen, other). After filling the questionnaire their results were sent to the server, and participants validated their participation on Mechanical Turk.

2.2.2 Materials

As in Bott and Noveck (2004, Exp. 3), participants were free to answer as they wished. The main difference is that we tested both direct and indirect implicatures.

There were 5 types of target sentences:

(D1) Some elephants are mammals.

(D2) John-the-zoologist believes that some elephants are mammals.

(I1) Not all elephants are reptiles.

(I2a) John-the-zoologist believes that not all elephants are reptiles.

(I2b) John-the-zoologist doesn't believe that all elephants are reptiles.

D1 are direct targets identical to those of Bott and Noveck (2004). They are true under the weak reading: "There are elephants which are mammals", and false under the strong reading (with implicature): "Some but not all elephants are mammals". I1 are indirect targets following the same pattern: true under the weak reading: "It is not the case that all elephants are reptiles", and false under the strong one: "Not all elephants are reptiles, but some are". It has been argued that the effect of Bott and Noveck may be due to their sentences being weird, because the inferences violate world knowledge[3]. Therefore, we decided to test different types of targets, in both direct (D2) and indirect (I2a, I2b) conditions. These sentences share the same properties as the simpler version: They are all true under the weak reading and false under the strong one. Unlike D1 and I1, their SIs do not violate world knowledge, but the instructions made clear that zoologists know everything there is to know about animals. For the indirect condition there were two possible positions for the negation which yielded the same truth conditions for the sentence and the SI (I2a and I2b). Both were tested. We generated 15 exemplars of each of these sentences by varying the animals and the names.

Control sentences were designed to ensure that participants could not guess the truth value of a sentence until reaching its end. For each target sentence one true control and one false control were built by changing the last word, as in the examples below:

(D1-T) Some elephants are Asian.

(D1-F) Some elephants are reptiles.

(I1-T) Not all elephants are Asian.

(I1-F) Not all elephants are mammals.

Controls for (D2) followed the model of (D1). Controls for (I2a – I2b) followed the model of (I1). As for targets, there were 15 exemplars of each of these sentences.

A set of 40 true affirmative fillers sentences were included to balance the extra negative sentences due to (I2b) and to push participants to give more false responses to targets by contrast (this turned out necessary from the results of a pilot study). They were not included in the analyses.

(Fill1) Some mammals are elephants.

[3]Magri (2009) showed how sentences are unnatural when the strong alternative is contextually equivalent to the weak one, as in this example: #Some Italians come from a warm country.

(Fill2) All elephants are mammals.

Overall, there were 75 target sentences, 150 control sentences, and 40 fillers. All sentences were displayed in blocks of 2-3 words (e.g., [John-the-zoologist][believes that][some elephants][are mammals]). Each block was displayed for 750ms, except for the last one which was displayed until a response was made.

2.2.3 Participants

42 participants were recruited and received $2.5 for their participation. 6 of them were removed from the analyses: one for not being a native speaker of English and five because their error rate to control sentences exceeded mean+sd: 18%. The mean error rate was 6% on the remaining participants (sd: 5%).

2.3 Results

2.3.1 Data treatment and statistical methods

Responses made in less than 100ms or more than 10s were removed from the analyses (1.8% of the data). Categorical responses were analyzed with mixed logit models (GLMM)[4]. For response time analyses, we removed all error trials. Log transformation was applied before fitting the models in order to respect homoscedasticity assumptions. All mixed models were built with the maximal structure for random effects as defined in Barr et al. (2013), with Subjects and Items as random factors. The details about random effects and the variance-covariance matrices are given in Appendix E. The intercepts are usually not meaningful, so we will not comment them. p-values are calculated from Wald's z-values for logit models and by treating t-values as z-values for linear models. For models on logRT, we sometimes give estimates of the size of the effects in milliseconds by applying an exponential transformation to the average fitted values. The error bars on all graphs correspond to standard errors of the mean.

2.3.2 Analysis of responses

Figure 1 shows the proportion of *True* answers given to controls and targets, aggregated by participants.

Controls: Responses to control sentences (correct vs. error) were fitted in a model including Truth value, Sentence type (D1, D2, I1, I2a, I2b) and their interaction as fixed effects. We observed a trend for an effect of Truth value on error rates: False sentences tend to yield more errors than true sentences ($z = 1.7$, $p = .09$). We observed no difference between D1 and D2; I1 ($z = -1.2; 0.6$, $p = .2; .5$, respectively), but significant differences between D1 and I2a; I2b($z = 2.4; 4.4$, $p = .02; < .001$, respectively): these sentences yielded more errors than affirmative ones. No interaction was significant. Replacing Sentence type with a simpler 2-level factor affirmative/negative and removing the interaction did not significantly reduce the explanatory power of the model: $\chi^2(7) = 10.6, p = .16$. In this case, both effects are significant: False sentences give rise to more errors ($z = -4.1$, $p < .001$) and so do Negative sentences ($z = -5.4$, $p < .001$).

[4]See Jaeger (2008) for strong arguments against the use of ANOVAs for categorial data analysis.

Targets: On targets the truth value depends on the participant's reading, whereas controls were unequivocally true or false. On average, participants gave 36% *False* responses to targets, which means they derived SIs 36% of the time, with little variation across the types of targets. A model was fit on this data (*True* vs. *False* responses): The fixed effects were Truth value (this time with 3 levels: target, true, false), Sentence type (D1–I2b) and their interaction. We found strong differences between the targets and both true and false controls: Target sentences gave rise to less *True* answers than true controls ($z = 5.5, p < .001$) but more than false controls ($z = -10, p < .001$). The proportion of *True* answers to targets was lower on I1 ($t = -2.3$, $p = .02$) and I2a targets ($t = -2.8$, $p = .005$), which means that participants derived more SIs in these cases, but the absolute percentages were still close (less than 10% difference). The interactions between Sentence Type and Truth value mostly represent the differences in error rates across sentences, which have already been commented on and are of little interest.

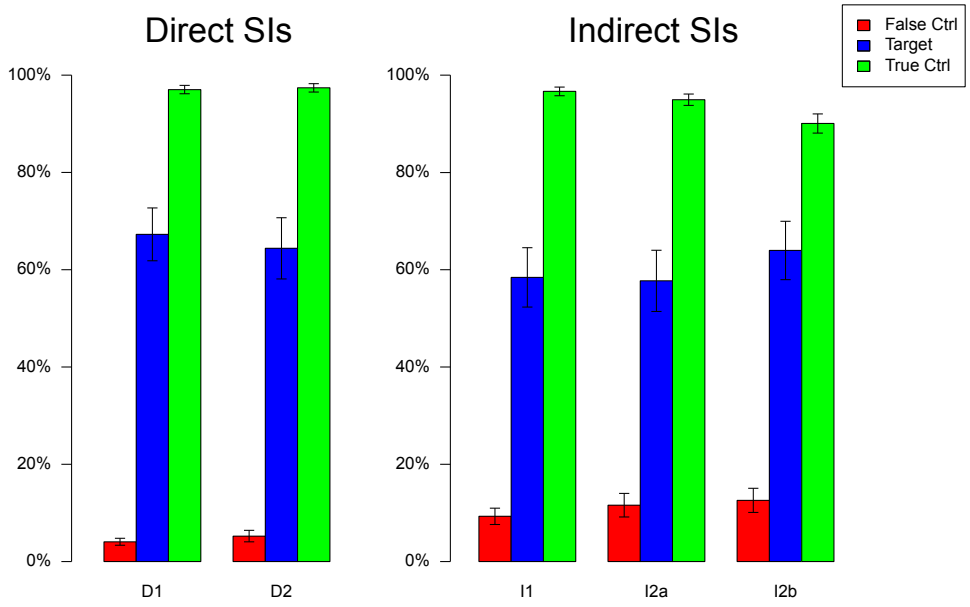


Figure 1: Percent *True* answers (aggregated by participant) as a function of Sentence Type.

We computed coherence indices by participant: absolute value of the difference between the derivation rates on direct and indirect targets and 50%. The median value was 38% and 34%, respectively, meaning that half the participants gave the same response at least 88% of the time to direct targets and at least 84% of the time for indirect targets. Therefore, many participants had a coherent behavior on targets and almost did not change readings throughout the experiment.

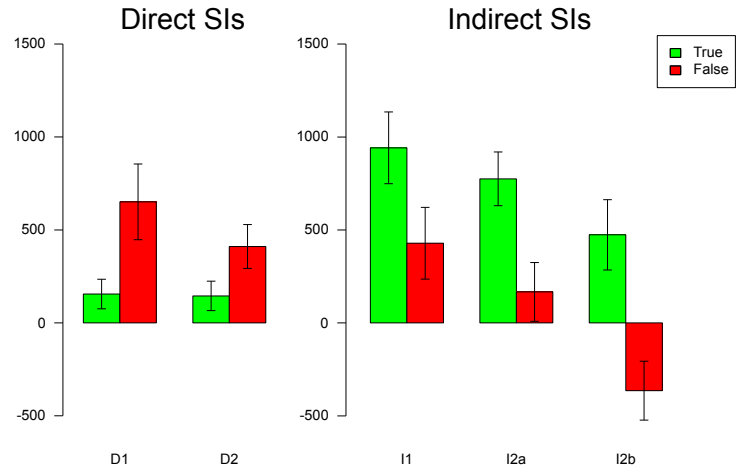
Finally we looked at the correlations between responses to the different targets. We observed a strong correlation between a participant's derivation rate (rate of *False* responses) for direct SIs and this participant's derivation rate for indirect SIs ($F(1, 34) = 24, p < .001$). All correlations between subtypes of targets (D1–I2b) were significant as well (all F's above

11, p-values under .001).[5]

2.3.3 Analysis of response times

Controls: A mixed model was fit on logRT from control sentences. The fixed effects were Truth value (true/false) and Sentence types (D1–I2b). We observed an effect of truth value: Responses to false sentences are slower by 109ms ($t = 2.9$, $p = .004$). We also observed that D1 sentences are significantly faster than all others (D2: +101ms, I1: +182ms, I2a: +304ms, I2b: +622ms, all $t > 2.2$, all $p < .05$). Truth value was not interacting with Sentence type except for a trend for I2a: The true/false difference in this case seems to be increased by an extra 214ms ($t = 1.8$, $p = .07$).

Targets: Figure 2a displays the fitted values for targets, from a mixed model which was fit on all data (controls and targets). The fixed effects were Answer (*True* vs. *False*), Control vs. Target, Sentence type (D1–I2b) and all interactions between these factors. For D1 sentences, answering a target does not take significantly longer than a control ($t = .09$, $p = .3$), but it tends to be the case on I1 and I2a sentences (441ms and 371ms, both $t = 1.7$ and $p = .09$ for interaction). Answering false to a D1 target has an extra cost, compared to answering false to a D1 control (+430ms, $t = 2.7$, $p = .008$). This corresponds to the effect discovered by Bott and Noveck (2004). However, this effect interacts with sentence type. Table 2b shows the estimated cost for deriving an SI in each sentence type. The difference between D1 and D2 is not significant ($t = -.6$, $p = .5$), but the effect is reversed (or tends to be) on negative sentences I1, I2a, I2b ($t = -1.6; -2.4; -2.8$, $p = .10; .02; .006$, respectively).



(a) RT difference between targets and controls. (aggregated by participant)

Type	Effect size
D1	430ms
D2	254ms
I1	-424ms
I2a	-655ms
I2b	-510ms

(b) Estimated size of the interaction between *True*/*False* effect and Target/Control sentence.

Figure 2: On direct targets, SIs seem costly (Bott and Noveck effect) but this effect is reversed on indirect targets.

[5] All these correlations were calculated on arcsine-transformed derivation rates. We did not use logit transformation because many participants had 0% or 100% rates.

2.4 Discussion

Our results first reveal a strong correlation between the derivation rates of direct and indirect SIs across participants. At the 'offline' level, we thus confirm expectations that the two types of inferences are two sides of the same coin.

For direct SIs we replicated the effect of Bott and Noveck (2004): SIs incur an extra cost which is not explained by true/false effects on controls. We observed no difference between their original sentences (D1) and those which did not violate world knowledge (D2). However, on indirect targets (I1, I2a-b), the effect was reversed. In negated sentences, the weak reading seems to be accessed later than the strong one. This result is very surprising. Indeed, if indirect SIs display a different processing signature in the paradigm of Bott and Noveck (2004), two conclusions come to mind. Either the theory must be revised, so that indirect SIs would not be explained by the same mechanism as direct ones, or we must question what the paradigm is actually testing. In the present case, we will argue that this experiment does not challenge the similarity between the two phenomena but rather can be accounted for by appealing to the nature of the control sentences we used.

First of all, the controls we used in this experiment shared their structure with targets. Since most participants were very coherent on targets and stuck to one reading (either weak or strong), if a participant was deriving SIs on the targets, he probably did so on controls as well, even though this does not affect truth-conditions. In this case it is not very clear what these sentences actually control for. If SIs have a derivation cost, it probably affected control sentences as well. The comparison between controls and targets would only yield information on the differences in semantic complexity, but would not tell much about derivation.

Second, there may be an asymmetry between direct and indirect targets. All sentences associated the name of an animal (e.g., "elephant") with a category that could either be a match ("mammals"), a mismatch ("reptiles") or a partial match ("Asian"). D1 involves a match and I1 involves a mismatch. This may have facilitated the *True* response to D1 but hindered it for I1. On true controls, the situation is symmetric as they both involved partial matches (D1-T is "Some elephants are Asian", I1-T is "Not all elephants are Asian"). For participants with a strong reading, the situation is reversed: If mismatches facilitate *False* responses, then D1 should be harder than I1. Therefore the double interaction we observed (True/False $\times$ Control/Target $\times$ D1/I1) may only reflect this asymmetry.

We observed slower responses to control sentences I1–I2b overall, and the false controls in particular could be slower (significant on I2a). This extra delay on false negative sentences is unexpected if we only consider effects of truth value and polarity. Classic experiments on the comprehension of negation (Clark and Chase, 1972; Carpenter and Just, 1975) show that if there is an interaction between these effects, then in negative sentences falsity is usually easier to verify than truth. If anything, our control sentences show the opposite pattern. Some specificities of these sentences (e.g. their own SIs, or their match/mismatch properties) may explain the surprising response times pattern we observed.

Therefore, we cannot conclude much about the processing of SIs from these results. It confirms that verification plays an important role in the results and the cost measured by previous sentence verification studies may reflect the impact of the inference on the

verification process rather than its derivation. The proper way to control for polarity and truth value effects would be to use controls which do not trigger SIs themselves, so that participants' behavior would affect their response times only on targets, and to have a wider variety of items in order to avoid simple patterns like the match/mismatch we observed here.

3 Experiment 2: controlling for superficial differences

3.1 Goal of this experiment

This experiment included additional control sentences, designed to solve the main issues we identified in Experiment 1. To avoid the match/mismatch issue discussed above, we used a cover-story so that the task would depend as little as possible on world knowledge and established categories. Such a move has already been done for independent motivations in Chemla and Bott (2011). We also greatly increased the variety of control sentences in order to avoid any pattern or prediction based on a subpart of the sentence. We included controls which did not trigger implicatures (with quantifiers 'all' and "none"), to be able to properly control for true/false and affirmative/negative effects. Finally, we moved to a design with training, as in Bott and Noveck (2004, Exp. 1): Participants were trained to rely either on the strong or weak reading to evaluate target sentences. This let us see how participants' behavior on the targets (deriving or not the implicature) may facilitate or hinder their responses to some controls. It also ensured an equal share between weak and strong responses, thus increasing statistical power for response times analyses. The first experiment already provided enough data on the derivation rates of direct and indirect SIs and established their strong correlation.

3.2 Methods and Materials

The experiment was a sentence-verification task, with a paradigm similar to Bott and Noveck (2004, Exp. 1): Participants were separated into 2 groups. Each group received specific instructions and training to use either weak or strong reading. We will refer to the weak reading group as *No-SI-group* and to those in the strong reading group as *SI-group*[6,7].

3.2.1 Course of the experiment

The course of this experiment was similar to Experiment 1. However, after being recruited on Mechanical Turk, participants were randomly assigned to the SI or No-SI group. The instructions began with a story participants had to read. The experimental sentences had to be judged against this story. It was identical in both versions. The instructions differed on the way the example-sentences were treated (cf. section 3.2.2 for details).

The experiment consisted in a training phase and an experimental phase. The training

[6]"No-SI" and "SI" correspond to Bott and Noveck's "literal" and "pragmatic" respectively. The nature of SIs is debated in the literature (pragmatic or grammatical). Our experiment and its results are independent from this question, and we therefore avoid the term "pragmatic" which is marked with respect to this debate, hence we will adopt more neutral terms.

[7]In Bott and Noveck (2004), each participant took the experiment in both a no-SI condition and an SI condition, allowing for within participant comparison. This would have made our experiment too long, due to the fact that we tested more cases of SIs.

was longer than in Experiment 1 (32 items), because participants had to learn to treat target sentences as true or false, depending on their group.

Between the training and the experimental phases, a summary of the story was presented. Within each phase (training and experimental) items order was randomized. Upon reaching half of the experimental phase, participants were offered to take a short break. During this break a reminder of the story presented in the instructions was displayed.

3.2.2 Materials

In the instructions, subjects read the following story:

> A new virus has emerged in a zoo. In just a few days all animals were infected. All land animals died after two weeks. On the contrary, birds were fortified by this unknown virus. Meanwhile, the vets who are working in the zoo inspected half of the animals from each species (they did not have the time to inspect them all but they wanted to see some animals from each species).

The most important aspects of this cover-story concern the proportion of each kind of animals which were inspected, killed or fortified. They were as follow:

	Land animals	Birds
Inspected	Half	Half
Killed	All	None
Fortified	None	All

The sentences used in the experiment followed the pattern: "QUANTIFIER of the ANIMALS were VERB". There were 4 possible quantifiers: 'some', 'all', "Not all" and "None". The animal name was taken from a list of 36 birds or from a list of 36 land animals (plus 8 of each for the training phase). The verbs were "killed", "fortified", and "inspected". Overall there were 24 types of sentences ($4 \times 2 \times 3$). The quantifiers 'some' and "not all" could give rise to SIs. The sentences (7a-b) and (8a-b) were direct and indirect targets respectively (all are true under the weak reading but false under the strong reading).

(7) a. Some of the [land animals] were killed.
    b. Some of the [birds] were fortified.

(8) a. Not all of the [land animals] were fortified.
    b. Not all of the [birds] were killed.

All of the 20 possible control sentences were included but each of them was assigned a different frequency, so that there would be an equal number of true and false sentences overall, for each quantifier, for each verb and for each type of animal. This was done to prevent subjects from developing strategies to predict the truth value of a sentence from one of its constituents, hence forcing them to pay attention to all words in the sentence. Quantifier 'all' provided the affirmative controls without implicature (true and false), while "none" provided the negative ones.

There was a total of 264 sentences, including 48 targets (24 direct - 24 indirect)[8]. The training phase was balanced in a similar way but had a higher proportion of target sentences (8 from a total of 32 items). The detailed list with the exact number of repetitions for each sentence type is reported in Appendix, as well as the animal lists.

In the instructions, subjects were presented 4 examples of sentences taken from the training phase. The first 2 were true and false controls, while the 2 others were direct and indirect target sentences. The instructions for these depended on the subject's group. During the training phase, subjects received feedback every time they made an error. On targets this feedback depended on the training as well.

Sentences were displayed word-by-word and each word was displayed for 250ms, except the last one on which we measured the response time.

3.2.3 Participants

84 participants were recruited on Amazon's Mechanical Turk. They received $2 for their participation. We removed 24 participants: 9 who were not native speaker of English, 4 with error rates to controls under mean plus one standard deviation (19.7%) and 11 with error rates under 40% on both direct and indirect targets. Overall, 27% of the participants were removed[9].

The mean error rate for remaining participants was 5.8% on controls and 4.9% on targets. From all removed participants, 11 out of 24 were in the SI training group, and from the 11 we removed specifically for their errors on targets, 5 were in the SI group and 6 in the no-SI group. A Welch Two Sample t-test showed no significant difference in the error rates on targets between the two training groups when only the non-native speakers of English were removed ($t(69) = 0.47$). A more detailed analysis is provided below.

3.3 Results

3.3.1 Data treatment and statistical methods

As in Experiment 1, responses made in less than 100ms or more than 10s were removed from the analyses (1.7% of the data). The analyses follow the same rules as in Experiment 1 (mixed logit models for categorial data and linear mixed models on response times after log transformation). All mixed models have a maximal random effect structure in the sense of Barr et al. (2013) and the details about the random effects can be found in appendix E. The error bars on all graphs correspond to standard errors of the mean.

[8]Note that there is a trade-off between a high number of repetitions for targets, and the size of the true/false bias that is due to these targets: Since we used the same controls for both groups of subjects, they all had 108 true and 108 false control sentences, but the 48 targets were all true for subjects from the no-SI group and all false for subjects from the SI group. We kept 24 direct targets and 24 indirect targets, yielding enough repetitions and limiting the bias to a 59% / 41% (also keeping the experiment not too long).

[9]One may think this proportion is very high, but there is a trade-off between keeping many participants and having a reasonable error rate on targets. We considered that when a participant had more than 40% error rate on one type of target the training had failed (probably because the participant had a very strong bias for the other reading and would not change it).

3.3.2 Analysis of responses

Figure 3a shows the proportion of *True* responses to all types of sentences. Accuracy was above 80% in all conditions.

Control sentences: We verified that participants understood the task and that training did not have unexpected effects on controls sentences.

A GLMM was fit on responses to controls (correct/incorrect) with the following fixed effects: Truth Value (True vs. False), Sentence Polarity (Affirmative vs. Negative) and Training (No-SI vs. SI), plus all possible interactions. On affirmative sentences, truth value had a significant effect ($z = 2.0, p = .43$). Negative sentences gave rise to more errors ($z = 4.1, p < .001$) and there was a significant interaction with truth value ($z = -3.0, p = .002$): Among negative sentences, false sentences give rise to less errors than true sentences. We also detected an interaction between Truth Value and Training: SI participants make less errors on False controls ($z = -3.0, p = .002$). No other effect was significant (all $|z| < 1$, $p > .4$).

Target sentences: Before removing participants who had too high error rate on targets or controls, we tested whether the training had the expected effect on responses to target sentences. It is necessary to check that it is the case even when participants who did not respond to targets as they had been trained are included in the analyses, otherwise the results would only reflect the way we selected the participants.

No-SI participants were trained to answer *True* to targets whereas SI participants were trained to answer *False*. We ran a GLMM on responses (*True*/*False*) with the following fixed effects: Polarity (Affirmative/Negative, which parallels Direct/Indirect for targets) and Training (No-SI/SI) plus their interaction and the interaction between Training and Sentence Type (Target/True control/False control)[10]. The model output is presented in figure 3b.

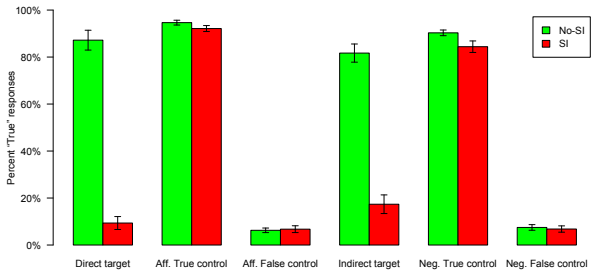
We observed a large effect of training on responses to targets. No-SI participants gave more *True* responses than SI participants (87% vs. 14%). Participants from both groups gave less *True* responses on negative targets (83% and 12% respectively). There may be a slightly stronger bias toward the strong reading in negative sentences, or a bias for more errors on negative sentences. The comparison with controls reveals that the difference between targets and false controls for no-SI participants or true controls for SI participants is significant (both $|z| > 13$).

3.3.3 Analysis of response times

We first removed all error trials (6.0% of the data).

Raw response times:

[10]Sentence Type was not treated as a simple effect. The formula for this model is a bit complicated: Answer∼Training*Polarity+Training×SentenceType. The point was to avoid treating No-SI×(Target→TrueCtrl) as a simple effect, and to treat it as an interaction instead. Thereby, it is not taken into account when computing SI×(Target→TrueCtrl). The latter is directly added to the effect of Training. The same holds for (Target→FalseCtrl). Considering SentenceType as a simple effect would have been conceptually equivalent, but the effects from this model are easier to interpret.



(a) Percent *True* responses by Training

Effect	β	z-value	p-value
Intercept	2.6	6.7	.00
Training: No-SI→SI	-5.7	-10.6	<.001 ***
Polarity: Dir → Indir	-0.3	-2.6	.01 *
Training × Polarity	-0.2	-1.0	.31
No-SI × (Target→TrueCtrl)	0.4	0.9	.35
No-SI × (Target→FalseCtrl)	5.6	13.4	<.001 ***
SI × (Target→TrueCtrl)	-5.4	-13.9	<.001 ***
SI × (Target→FalseCtrl)	0.1	0.2	.81

(b) GLMM on responses (*True*/*False*). Intercept corresponds to the proportion of *True* responses given by No-SI participants to affirmative (direct) targets.

Figure 3: No-SI participants treat targets as true sentences while SI participants treat them as false sentences. Participants from both groups give very similar responses to controls.

Control sentences: A mixed model with 3 fixed factors (Participants training, Polarity and Truth Value) was fit on the response times from control sentences. To correct for heteroscedasticity, a log transformation was applied to the RTs.

We observed an effect of Truth Value (False sentences take about 68ms more than True ones, $t = 3.2$, $p = .001$) and Polarity (Negative sentences take about 281ms more $t = 7.1$, $p < .001$). No other effect was significant (all $|z| < 1.1$, $p > .2$). The training had no effect on RT for controls and crucially did not interact with Polarity.

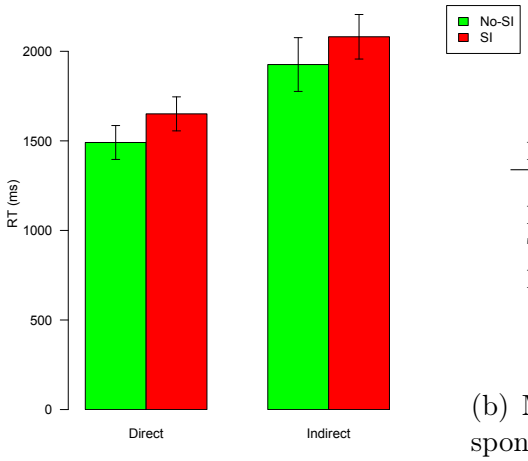
Target sentences: Figure 4a shows the response times for target sentences. A mixed model was fit on these log response times before any further treatment. The fixed effects were Polarity (Direct/Indirect), Participant training, and their interaction. The model output is presented in table 4b. As on controls, we observed a significant effect of Polarity (Indirect slower than Direct by 345ms). As in Bott and Noveck (2004), we observed an effect of the training (SI participants are overall slower on targets by 194ms). This time there was no interaction between Training and Polarity.

However, since Training parallels the answer given by the participants (No-SI participants had to answer *True*, whereas SI participants had to answer *False*), it could just be a true/false effect. We also observed an effect of polarity on controls, which may explain the difference between direct and indirect targets.

Once again, these results are not sufficient to reach a conclusion. Without controlling for Polarity and Truth Value, it is hard to know how much of the effect is due to superficial properties of the target sentences, and how much comes from the derivation of the SI.

3.3.4 Regressing the effects of Polarity and Truth value

As explained earlier, the main feature of Bott and Noveck's paradigm is to associate a response (*True/False*) with a given reading (Weak/Strong). Because of this we must control very carefully for true/false effects. In our experiment, we also have to disentangle the effect of Polarity (affirmative/negative) from the type of implicature (direct/indirect). Furthermore, Polarity and Truth value usually interact in sentence verification tasks, although it



(a) RT on targets

Effect	β	t	p
(Intercept)	7.129	141	0
Polarity: Dir.→Indir.	0.244	5.8	<.001 ***
Training: No-SI→SI	0.144	2.1	.036 *
Polarity×Training	-0.071	-.4	.7

(b) Mixed model on RT for targets. Intercept corresponds to responses given by no-SI participants to direct targets.

Figure 4: We replicate Bott and Noveck (2004) effect: On targets, no-SI participants are faster than SI participants. This effect does not interact with Polarity.

does not seem to be the case here.

We used the response times measured on control sentences to regress the effects of Truth value, Polarity and their interaction from target sentences[11]. Details about the procedure for residual response times (rRTs) are provided in Appendix D. rRTs for targets are presented in Figure 5a. They were analyzed with a linear mixed model which output is presented in table 5b[12].

The previous effect of Polarity disappeared ($p = .19$ instead of $p < .001$), while the effect of Training *à la* Bott and Noveck (2004) remained statistically visible ($p = .10$ instead of $p = .036$). Crucially, there was no interaction between the effects of Training and Polarity.

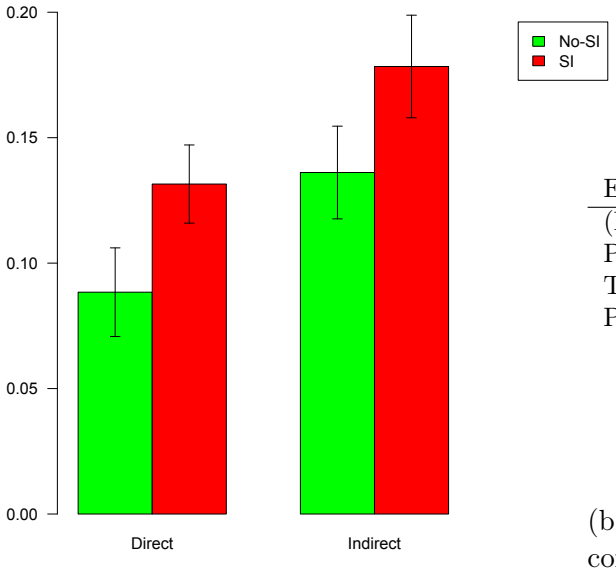
The cost of deriving the SI can be estimated by looking at the data after taking the exponential of rRT and multiplying them by the ratio of their standard deviation to the standard deviation for raw RT. We obtain an estimation of 99ms, which is compatible with the results of Bott et al. (2012), who showed that the cost associated with the SI is about 100ms once the effects of sentence verification are removed (using a SAT design).

3.4 Discussion

On control sentences, we replicate the classic results about sentence comprehension and negation (Clark and Chase, 1972; Carpenter and Just, 1975; Kaup et al., 2006, 2007). Falsity and negation delay the comprehension of the sentence, but they may interact negatively: False negative sentences are not as hard as expected. We did not observe a significant

[11]DeGutis et al. (2012) showed that regression is a more efficient way to control for an effect than subtraction. Conceptually, subtracting is equivalent to a regression in which the offset is fixed to zero and the slope to 1. Letting a model choose the appropriate values is the proper way to remove all the variance in targets which can be explained by the controls.

[12]There was no need for a log-transformation since rRT are the output of a model on logRT. A visual inspection of the residuals confirmed the homoscedasticity hypohtesis.



(a) Residual RT (aggregated by participant)

Effect	β	t-value	p-value
(Intercept)	8.4	3.5	0
Polarity: Dir.→Indir.	5.3	1.3	0.19
Training: No-SI→SI	5.2	1.6	0.10 (.)
Polarity×Training	-1.0	-0.2	0.85

(b) Linear mixed model on rRT. Intercept corresponds to responses given by no-SI participants to direct targets.

Figure 5: Bott and Noveck (2004) effect (No-SI participants faster than SI) is present as a trend. This effect does not interact with Polarity: Direct and Indirect SIs do not differ.

interaction, but it would go in this direction if anything. This is a crucial difference between Experiments 1 and 2.

On targets, the results on raw RT replicate the effect of Bott and Noveck (2004) in affirmative sentences (direct SIs). However, this effect may be related to the true/false effect we observed on controls. Furthermore, direct and indirect SIs sentences differ in polarity. To disentangle these two possible confounds we controlled for polarity, truth value and their interaction. After regression, the interaction is still absent whereas the main effect of training (SI vs. No-SI) does not disappear.

As a conclusion, our results suggest that there is no real difference between direct and indirect implicatures with regard to the derivation cost, although this fact may be hidden by superficial differences. The differences between Experiments 1 and 2 show how control sentences must be chosen and used very carefully.

4 General discussion

4.1 Summary of the results

Our goal was to compare direct and indirect SIs. Our experiments provided both offline and online data.

On the offline side, in Experiment 1 we did not observe big differences between the derivation rates of direct and indirect SIs. Furthermore, the derivation of these two types of inferences strongly correlated within subjects.

On the online side, the picture looks slightly more complicated at first. In Experiment 1, indirect implicatures seemed to have a different processing signature from direct implicatures.

However, we suggested that this difference could be an artifact due to (i) specific properties of the controls, which may also trigger SIs, and (ii) a superficial difference between our positive and negative target sentences, which involved either a match or a mismatch between the animals and animal species mentioned. In Experiment 2 we showed that the difference between direct and indirect SIs does fade away with appropriate controls. The two types of implicatures showed similar processing patterns, even after controlling for the two possible confounds: polarity and truth value.

Our results are coherent with the broad consensus in the theoretical literature that direct and indirect SIs result from the same processes.

4.2 Bott and Noveck's paradigm provides a real signature of scalar implicatures

As shown by the differences between Experiments 1 and 2, Bott and Noveck's paradigm is very sensitive to the properties of the control sentences. By associating one response (*True* or *False*) with a behavior (e.g., deriving or not deriving an SI), the effect related to the phenomenon at stake are likely to be mixed with unwanted effects of truth value. This is even more problematic when elements that interact with truth value effects play a central role, in the way negation distinguishes between direct and indirect SIs.

However, Experiment 2 shows that when True/False effects are properly controlled for, the paradigm provides a reliable processing signature of scalar implicatures. This processing signature can thus serve two lines of investigations. First it can be used to compare SIs with other phenomena, such as presuppositions (Chemla and Bott, 2011) and free-choice inferences (Chemla and Bott, 2012). Second, the processing cost can be linked to the SI mechanisms more firmly. A closer investigation of the possible sources for this cost is therefore required, which discussion we move to in the next paragraph.

4.3 About the cost of scalar implicatures

As we explained in the introduction, the experimental literature on scalar implicatures is abundant and most studies detected a cost for deriving an implicature. Bott et al. (2012) used a SAT paradigm which allowed them to distinguish the delays coming from semantic complexity and from derivation processes. They proved that a cost of about 140ms is specifically due to the derivation process. Our estimates for direct and indirect implicatures (about 100ms) are compatible with their results.

Hence, the derivation process of SIs is costly, and this cost is produced equally in the derivation of direct and indirect SIs. But it is not clear what in the SI derivation process creates this cost. All derivation processes of SIs describe additional operations to obtain an implicature (see Chemla and Singh 2013) and a cost could be associated to any of these operations. For instance, it could be associated with (a) the mere decision to derive an implicature, (b) with the retrieval of the necessary alternatives, (c) with the comparison between the utterance and its alternatives, (d) with the negation of the selected alternatives, etc.

A study of the contribution of each of these sources seems possible and important. For instance, we could study how adding more scalar alternatives (e.g., including "most" as an alternative to 'some') could impact the processing of scalar implicatures. Degen and Tanen-

haus (2011) already showed that 'some' was processed more slowly when new alternatives are made relevant, independently of the derivation of an implicature. If the addition of alternatives also delays the derivation of the implicature, this would suggest that (b), (c) or (d) above could participate to the effect. On the other hand, De Neys and Schaeken (2007) showed that increasing the working memory load reduces the rate at which participants derive scalar implicatures, but Marty and Chemla (2013) further observed no such effect on explicit paraphrases with 'only', which involve the same manipulation of alternatives. This may suggest that the processing of the alternatives is not the most effortful part in the derivation of implicatures and therefore argues in favor of (a) in the options above. In any event, more investigations are needed to understand this processing cost. The mechanisms at the source of direct and indirect SIs trigger this cost, while other similar phenomena do not. The different operations postulated in formal derivation models of SIs should now provide a set of hypotheses to identify the source of the cost.

5 Conclusion

In the last decade, profuse literature on the processing of scalar implicatures has provided strong arguments for a costly derivation. By eliminating confounds and extending previous results to a wider class of scalar implicatures, our results reinforce the idea that this cost is a core property of these inferences and may be enough to distinguish them from other phenomena. Yet formal theories of the phenomena postulate several steps in the derivation of implicatures and it is still unclear which of these steps are costly, therefore more research is required to reach a full understanding of this cost and bind formal, representational models with experimental, processing results.

References

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Bott, L., Bailey, T. M., and Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, 66(1):123 – 142.

Bott, L. and Noveck, I. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 53:437–457.

Carpenter, P. A. and Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, 82(1):45–73.

Chemla, E. (2009). Similarity: Towards a unified account of scalar implicatures, free choice permission, and presupposition projection. Under Revision for *Semantics and Pragmatics*.

Chemla, E. and Bott, L. (2011). Processing presuppositions: Processing presuppositions: dynamic semantics vs pragmatic enrichment. *Language and Cognitive Processes*.

Chemla, E. and Bott, L. (2012). Processing: free choice at no cost. In *Logic, Language and Meaning*, pages 143–149. Springer.

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

Chemla, E. and Singh, R. (2013). Remarks on the experimental turn in the study of scalar implicature. Under Revision for *Language and Linguistic Compass*.

Chemla, E. and Spector, B. (2011). Experimental evidence for embedded scalar implicatures. *Journal of Semantics*, 28:359–400.

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

De Neys, W. and Schaeken, W. (2007). When people are more logical under cognitive load. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie)*, 54(2):128–133.

Degen, J. and Tanenhaus, M. K. (2011). Making inferences: the case of scalar implicature processing. In *Proceedings of the 33rd annual conference of the Cognitive Science Society*, pages 3299–3304.

DeGutis, J., Wilmer, J., Mercado, R. J., and Cohan, S. (2012). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition*.

Fox, D. (2007). Free choice disjunction and the theory of scalar implicature. In Sauerland, U. and Stateva, P., editors, *Presupposition and Implicature in Compositional Semantics*, pages 71–120. Palgrave Macmillan, New York, NY.

Grice, H. (1967). Logic and conversation. William James Lectures, Harvard University.

Huang, Y. T. and Snedeker, J. (2009a). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Pyschology*, 58:376–415.

Huang, Y. T. and Snedeker, J. (2009b). Semantic and pragmatic interpretation in 5-year olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, 45:1723–1739.

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434–446.

Katzir, R. (2007). Structurally defined alternatives. *Linguistics and Philosophy*, 30:669–690.

Kaup, B., Lüdtke, J., and Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, 38(7):1033 – 1050.

Kaup, B., Yaxley, R. H., Madden, C. J., Zwaan, R. A., and Lüdtke, J. (2007). Experiential simulations of negated text information. *The Quarterly Journal of Experimental Psychology*, 60(7):976–990.

Kratzer, A. and Shimoyama, J. (2002). Indeterminate pronouns: The view from Japanese. In *3rd Tokyo Conference on Psycholinguistics*.

Magri, G. (2009). A theory of individual level predicates based on blind mandatory scalar implicatures. *Natural Language Semantics*, 17:245–297.

Marty, P. P. and Chemla, E. (2013). Scalar implicatures: working memory and a comparison with "only". *Frontiers in Psychology*, 4:403.

Noveck, I. A. and Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, 85(2):203–210.

Romoli, J. (2013). The presuppositions of soft triggers are obligatory scalar implicatures. *Journal of Semantics*.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, 27:367–391.

Spector, B. (2003). Scalar implicatures: Exhaustivity and gricean reasoning. In ten Cate, B., editor, *Proceedings of the eigth ESSLLI student session*, Vienna, Austria.

Spector, B. (2005). Scalar implicatures: Exhaustivity and gricean reasoning. In Aloni, M., Butler, A., and Dekker, P., editors, *Questions in Dynamic Semantics*. Elsevier, Amsterdam.

van Rooij, R. and Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language, and Information*, 13:491–519.

van Tiel, B., van Miltenburg, E., Zevakhina, N., and Geurts, B. (2013). Scalar diversity. Unpublished MS.

Appendix A: List of all item bases for Experiment 2

Repetitions	Type	Quantifier	Species	Verb
Targets				
12	Direct	Some of the	[birds] were	fortified
12		Some of the	[animals] were	killed
12	Indirect	Not all of the	[birds] were	killed
12		Not all of the	[animals] were	fortified
Controls – Same structure				
9	True affirmative	Some of the	[birds] were	checked
9		Some of the	[animals] were	checked
9	True negative	Not all of the	[birds] were	checked
9		Not all of the	[animals] were	checked
9	False affirmative	Some of the	[birds] were	killed
9		Some of the	[animals] were	fortified
9	False negative	Not all of the	[birds] were	fortified
9		Not all of the	[animals] were	killed
Controls – No implicature				
18	True affirmative	All of the	[birds] were	fortified
18		All of the	[animals] were	killed
18	True negative	None of the	[birds] were	killed
18		None of the	[animals] were	fortified
9	False affirmative	All of the	[birds] were	killed
9		All of the	[birds] were	checked
9		All of the	[animals] were	fortified
9		All of the	[animals] were	checked
9	False negative	None of the	[birds] were	checked
9		None of the	[birds] were	fortified
9		None of the	[animals] were	checked
9		None of the	[animals] were	killed

Appendix B: List of animals for Experiment 2

Nouns used for the training phase are written in italic.

Land animals: Antelopes, bears, boa constrictors, buffaloes, cats, chameleons, cheetahs, chimpanzees, deer, dogs, elephants, foxes, giraffes, gorillas, horses, hyenas, iguanas, jaguars, kangaroos, koalas, lions, llamas, mongooses, opossums, pandas, panthers, pigs, ponies, pythons, rabbits, rhinoceros, sloths, squirrels, tigers, wolves, zebras, *camels, donkeys, hares, leopards, moose, rabbits, skunks.*

Birds: Albatrosses, blackbirds, blue jays, canaries, cardinals, condors, cormorants, crows, cuckoos, dodos, ducks, eagles, falcons, flamingos, geese, gulls, hawks, hens, jackdaws, ostriches, owls, parrots, peacocks, pelicans, pheasants, pigeons, ravens, robins, seagulls, storks,

swallows, swans, toucans, turkeys, vultures, woodpeckers, *doves, finches, herons, hummingbirds, pheasants, roadrunners, shorebirds, thrushes.*

Appendix C: Specific instructions and feedback to bias subjects toward weak or strong reading

In the instructions, no-SI subjects were told that the target sentences were true, although they were not the best descriptions of the situation. SI subjects were told that these sentences were false because their stronger alternatives were true (with scalar item written uppercase). For instance (9b) shows instructions for the no-SI group for a direct example (9a), while (9c) shows the instructions for the SI group.

(9) a. Some of the hummingbirds were fortified.
   b. This is ***True***: All of the hummingbirds were fortified, therefore some of them were.
   c. This is ***False***: Hummingbirds are birds and ALL of the birds were fortified.

During the training, the feedback on target was also adapted to the subject's group, although all subjects received the same feedback on control sentences:

(10) Target sentence: Some of the hummingbirds were fortified.
   a. No-SI feedback: It was True! Hummingbirds are birds. In fact all of them were fortified.
   b. SI feedback: It was False! Hummingbirds are birds. ALL of them were fortified.

(11) Control sentence: All of the leopards were killed.
Feedback: It was True! Leopards are land animals. All of them were killed.

Appendix D: Regression of Polarity and Truth value

Since there is no one-to-one correspondence between controls and targets, it was not possible to run a simple regression. Instead we ran a model on log response times with Polarity, Truth value and their interaction as fixed effects. We gave weight 0 to targets and 1 to controls. With such weights, the coefficients of the model are computed on control sentences only, but the model extrapolates its prediction to target sentences. Therefore, the residuals for targets correspond to the target data, from which the effects of polarity and truth value measured on controls have been removed. The random effect structure was maximal, with Subject and Item as random effect, so that the regression would also remove between-subjects and between-items differences. We defined residual response times (rRT) as the residuals from this model.

At this point, we trimmed the data by removing the first and last percentiles. We did not want to remove extreme percentiles before removing between-subjects differences because there were large differences between subjects and we would have removed a lot of data form the fastest subjects (although their error rates were usually low) instead of abnormal fast responses. We also did not want to cut the data beyond one or two standard deviations because it was still slightly skewed toward long RT, even after log-transformation. This

would mostly have resulted in a loss of data on sentences which trigger long responses, e.g. targets. By cutting 2sd from the mean, we would have lost 5% of the data, mostly on the high-end.

Appendix E: Variance-Covariance matrices for all mixed models

Information about the models are given in the same order as in the paper. For each model we provide the formula used in lmer or glmer (from R lme4 package) and the variance-covariance matrices for Subject and Item random effects.

In both experiments, the Item correspond to an animal name. This is the most careful definition, since two different sentences which involve the same animal name will be treated as non-independent measures (e.g. "Some elephants are mammals" and "Not all elephants are reptiles").

Experiment 1

`Correct~Value*SenType+(1+Value*SenType|Subject)+(1+Value*SenType|Item)`

Subjects

	(Intercept)	False	D2	I1	I2a	I2b	False:D2	False:I1	False:I2a	False:I2b
(Intercept)	1.625	-1.374	0.070	-0.362	-0.689	-0.748	0.532	1.040	1.813	1.910
False	-1.374	1.369	-0.139	0.323	0.414	0.418	-0.583	-0.931	-1.377	-1.411
D2	0.070	-0.139	0.348	-0.021	0.078	-0.126	-0.647	-0.054	-0.226	0.009
I1	-0.362	0.323	-0.021	0.398	0.377	0.302	-0.028	-0.822	-0.871	-0.531
I2a	-0.689	0.414	0.078	0.377	0.653	0.650	-0.134	-0.791	-1.248	-1.077
I2b	-0.748	0.418	-0.126	0.302	0.650	0.854	0.358	-0.539	-1.025	-1.169
False:D2	0.532	-0.583	-0.647	-0.028	-0.134	0.358	1.955	0.458	0.899	0.437
False:I1	1.040	-0.931	-0.054	-0.822	-0.791	-0.539	0.458	1.905	2.163	1.383
False:I2a	1.813	-1.377	-0.226	-0.871	-1.248	-1.025	0.899	2.163	3.096	2.459
False:I2b	1.910	-1.411	0.009	-0.531	-1.077	-1.169	0.437	1.383	2.459	2.499

Items

	(Intercept)	False	D2	I1	I2a	I2b	False:D2	False:I1	False:I2a	False:I2b
(Intercept)	0.159	-0.256	0.216	-0.243	-0.062	-0.038	0.008	0.251	0.125	0.110
False	-0.256	0.790	-0.458	0.416	0.066	0.116	0.272	-0.885	-0.505	-0.691
D2	0.216	-0.458	0.363	-0.368	-0.088	-0.125	-0.054	0.532	0.308	0.371
I1	-0.243	0.416	-0.368	0.451	0.108	0.099	-0.092	-0.467	-0.257	-0.241
I2a	-0.062	0.066	-0.088	0.108	0.033	0.030	-0.043	-0.074	-0.040	-0.020
I2b	-0.038	0.116	-0.125	0.099	0.030	0.111	0.025	-0.211	-0.154	-0.221
False:D2	0.008	0.272	-0.054	-0.092	-0.043	0.025	0.355	-0.310	-0.190	-0.377
False:I1	0.251	-0.885	0.532	-0.467	-0.074	-0.211	-0.310	1.082	0.656	0.928
False:I2a	0.125	-0.505	0.308	-0.257	-0.040	-0.154	-0.190	0.656	0.413	0.599
False:I2b	0.110	-0.691	0.371	-0.241	-0.020	-0.221	-0.377	0.928	0.599	0.930

`Answer~Value*SenType+(1+Value*SenType|Subject)+(1+Value*SenType|Item)`[13]

Subjects

	(Intercept)	False	True	D2	I1	I2a	I2b	False:D2	True:D2	False:I1	True:I1	False:I2a	True:I2a	False:I2b	True:I2b
(Intercept)	4.822	-4.729	-4.611	1.984	-0.999	-0.129	-1.413	-1.044	-2.636	2.109	0.672	1.084	0.340	1.630	0.432
False	-4.729	4.916	4.163	-2.128	1.124	0.407	1.713	1.077	2.987	-2.249	-0.655	-1.190	-0.199	-1.899	-0.294
True	-4.611	4.163	6.202	-1.996	0.681	-0.161	1.436	0.558	2.528	-2.373	-0.895	-2.001	-0.946	-2.715	-1.441
D2	1.984	-2.128	-1.996	1.158	-0.646	-0.562	-0.942	-0.309	-1.591	1.058	0.322	0.892	0.456	1.152	0.277
I1	-0.999	1.124	0.681	-0.646	3.314	2.939	3.391	-0.281	0.157	-3.234	-1.952	-2.755	-2.389	-3.516	-1.953
I2a	-0.129	0.407	-0.161	-0.562	2.939	3.661	3.878	-0.547	0.280	-2.705	-1.346	-3.146	-2.860	-4.363	-2.523
I2b	-1.413	1.713	1.436	-0.942	3.391	3.878	5.419	-0.370	0.920	-4.263	-1.872	-4.419	-3.105	-6.816	-3.776
False:D2	-1.044	1.077	0.558	-0.309	-0.281	-0.547	-0.370	0.784	0.847	0.248	0.280	0.787	0.665	0.811	0.536
True:D2	-2.636	2.987	2.528	-1.591	0.157	0.280	0.920	0.847	2.773	-0.833	0.111	-0.622	0.051	-1.194	-0.096
False:I1	2.109	-2.249	-2.373	1.058	-3.234	-2.705	-4.263	0.248	-0.833	4.270	2.212	3.729	2.352	5.362	2.790
True:I1	0.672	-0.655	-0.895	0.322	-1.952	-1.346	-1.872	0.280	0.111	2.212	1.439	1.760	1.277	2.182	1.206
False:I2a	1.084	-1.190	-2.001	0.892	-2.755	-3.146	-4.419	0.787	-0.622	3.729	1.760	4.248	2.958	6.100	3.437
True:I2a	0.340	-0.199	-0.946	0.456	-2.389	-2.860	-3.105	0.665	0.051	2.352	1.277	2.958	2.855	3.861	2.592
False:I2b	1.630	-1.899	-2.715	1.152	-3.516	-4.363	-6.816	0.811	-1.194	5.362	2.182	6.100	3.861	9.567	5.369
True:I2b	0.432	-0.294	-1.441	0.277	-1.953	-2.523	-3.776	0.536	-0.096	2.790	1.206	3.437	2.592	5.369	3.499

Items

[13] This model was further simplified but we kept the same random structure.

	(Intercept)	False	True	D2	I1	I2a	I2b	False:D2	True:D2	False:I1	True:I1	False:I2a	True:I2a	False:I2b	True:I2b
(Intercept)	0.191	-0.021	-0.219	0.065	-0.109	-0.063	-0.253	-0.144	0.028	-0.057	-0.065	-0.188	0.007	0.058	0.138
False	-0.021	0.405	0.153	-0.005	-0.153	-0.080	-0.080	0.036	0.205	-0.226	0.044	-0.185	0.105	-0.322	0.057
True	-0.219	0.153	0.451	-0.005	0.064	0.027	0.225	-0.145	0.175	0.078	-0.149	0.156	-0.069	-0.088	-0.186
D2	0.065	-0.005	-0.005	0.052	-0.042	-0.030	-0.099	-0.187	0.071	0.019	-0.124	-0.061	-0.039	0.052	0.025
I1	-0.109	-0.153	0.064	-0.042	0.185	0.070	0.133	0.196	-0.103	-0.007	0.064	0.171	-0.056	0.134	-0.100
I2	-0.063	-0.080	0.027	-0.030	0.070	0.086	0.146	0.133	-0.068	0.085	0.023	0.080	-0.074	0.005	-0.046
I3	-0.253	-0.080	0.225	-0.099	0.133	0.146	0.456	0.170	-0.117	0.272	0.074	0.302	-0.047	-0.078	-0.192
False:D2	-0.144	0.036	-0.145	-0.187	0.196	0.133	0.170	1.088	-0.313	-0.377	0.578	0.046	0.063	-0.164	0.037
True:D2	0.028	0.205	0.175	0.071	-0.103	-0.068	-0.117	-0.313	0.259	-0.059	-0.207	-0.133	-0.020	-0.096	0.005
False:I1	-0.057	-0.226	0.078	0.019	-0.007	0.085	0.272	-0.377	-0.059	0.520	-0.216	0.212	-0.111	0.096	-0.120
True:I1	-0.065	0.044	-0.149	-0.124	0.064	0.023	0.074	0.578	-0.207	-0.216	0.402	0.032	0.157	-0.126	0.041
False:I2a	-0.188	-0.185	0.156	-0.061	0.171	0.080	0.302	0.046	-0.133	0.212	0.032	0.312	-0.025	0.116	-0.181
True:I2a	0.007	0.105	-0.069	-0.039	-0.056	-0.074	-0.047	0.063	-0.020	-0.111	0.157	-0.025	0.182	-0.101	0.037
False:I2b	0.058	-0.322	-0.088	0.052	0.134	0.005	-0.078	-0.164	-0.096	0.096	-0.126	0.116	-0.101	0.370	-0.041
True:I2b	0.138	0.057	-0.186	0.025	-0.100	-0.046	-0.192	0.037	0.005	-0.120	0.041	-0.181	0.037	-0.041	0.134

`log(RT)~Value*SenType+(1+Value*SenType|Subject)+(1+Value*SenType|Item)`

Subjects

	(Intercept)	False	D2	I1	I2a	I2b	False:D2	False:I1	False:I2a	False:I2b
(Intercept)	0.091	-0.010	-0.003	0.008	0.006	0.005	0.002	0.002	-0.001	0.014
False	-0.010	0.021	0.006	0.004	0.008	0.001	-0.006	-0.016	0.000	0.003
D2	-0.003	0.006	0.012	0.013	0.016	0.022	-0.008	-0.012	0.001	-0.010
I1	0.008	0.004	0.013	0.021	0.026	0.038	-0.006	-0.017	0.003	-0.014
I2a	0.006	0.008	0.016	0.026	0.034	0.047	-0.005	-0.025	0.007	-0.015
I2b	0.005	0.001	0.022	0.038	0.047	0.086	-0.002	-0.027	0.021	-0.028
False:D2	0.002	-0.006	-0.008	-0.006	-0.005	-0.002	0.021	0.013	0.019	0.014
False:I1	0.002	-0.016	-0.012	-0.017	-0.025	-0.027	0.013	0.038	0.008	0.017
False:I2a	-0.001	0.000	0.001	0.003	0.007	0.021	0.019	0.008	0.067	0.035
False:I2b	0.014	0.003	-0.010	-0.014	-0.015	-0.028	0.014	0.017	0.035	0.073

Items

	(Intercept)	False	D2	I1	I2a	I2b	False:D2	False:I1	False:I2a	False:I2b
(Intercept)	0.007	-0.005	-0.012	-0.001	-0.001	0.005	0.004	-0.003	0.002	0.002
False	-0.005	0.021	0.005	0.008	0.021	0.022	0.003	0.008	-0.017	-0.021
D2	-0.012	0.005	0.190	-0.092	-0.033	0.005	0.034	0.249	0.094	0.026
I1	-0.001	0.008	-0.092	0.059	0.031	0.004	-0.021	-0.130	-0.056	-0.022
I2a	-0.001	0.021	-0.033	0.031	0.085	0.038	0.056	0.099	0.019	-0.029
I2b	0.005	0.022	0.005	0.004	0.038	9.186	-0.754	-0.609	0.490	0.866
False:D2	0.004	0.003	0.034	-0.021	0.056	-0.754	0.165	0.289	0.058	-0.079
False:I1	-0.003	0.008	0.249	-0.130	0.099	-0.609	0.289	0.788	0.257	-0.040
False:I2a	0.002	-0.017	0.094	-0.056	0.019	0.490	0.058	0.257	0.172	0.077
False:I2b	0.002	-0.021	0.026	-0.022	-0.029	0.866	-0.079	-0.040	0.077	0.118

`log(RT)~Answer*Target*SenType+(1+Answer*Target*SenType|Subject)+(1+Answer*Target*SenType|Item)` [14]

Subjects (with a factor 1,000)

	(Intercept)	False	Target	D2	I1	I2a	I2b	False Target	False D2	False I1	False I2a	False I2b	Target D2	Target I1	Target I2a	Target I2b	False Target D2	False Target I1	False Target I2a	False Target I2b
(Intercept)	1.064	-.003	-.056	-.059	-.028	-.060	-.100	.049	-.330	-.040	.074	.160	-.044	.291	.057	-.032	.189	.305	.459	.126
False	-.003	.177	.167	.156	.131	.039	.090	.116	-.008	-.197	-.081	-.011	-.145	-.096	-.065	-.246	-.011	.220	-.259	.031
Target	-.056	.167	1.376	.440	-.264	.321	.081	-.193	-.344	.254	-.030	.035	-.416	-.794	-.280	-.685	.320	-.080	.151	-.282
D2	-.059	.156	.440	.724	.109	.337	.305	.481	-.322	-.042	-.070	.085	-.370	-.143	.089	-.178	-.096	.199	-.505	-.307
I1	-.028	.131	-.264	.109	.915	.136	.428	.442	-.187	-.454	-.006	-.143	-.125	-.324	.021	-.015	-.042	.010	-.272	-.161
I2a	-.060	.039	.321	.337	.136	.622	.266	.056	-.303	-.023	-.438	-.196	.035	-.466	-.503	-.056	-.254	-.541	-.234	-.365
I2b	-.100	.090	.081	.305	.428	.266	.782	.291	-.244	-.350	.179	-.328	-.050	-.214	.206	-.363	-.202	.020	-.442	-.095
False:Target	.049	.116	-.193	.481	.442	.056	.291	1.147	-.387	-.279	.134	.232	-.265	-.085	.283	-.084	-.285	.325	-.635	-.439
False:D2	-.330	-.008	-.344	-.322	-.187	-.303	-.244	-.387	1.648	.204	.051	.159	.118	.337	-.286	-.205	-1.101	.068	-.222	.409
False:I1	-.040	-.197	.254	-.042	-.454	-.023	-.350	-.279	.204	.975	.316	.467	.099	.050	-.134	.011	-.104	-.259	.945	.427
False:I2a	.074	-.081	-.030	-.070	-.006	-.438	.179	.134	.051	.316	1.747	.609	-.258	.328	1.130	-.329	.245	.824	.408	.582
False:I2b	.160	-.011	.035	.085	-.143	-.196	-.328	.232	.159	.467	.609	1.221	-.176	.117	.301	-.203	.014	.507	.250	.146
Target:D2	-.044	-.145	-.416	-.370	-.125	.035	-.050	-.265	.118	.099	-.258	-.176	.938	.158	.024	.186	-.740	-.492	.401	.438
Target:I1	.291	-.096	-.794	-.143	-.324	-.466	-.214	-.085	.337	.050	.328	.117	.158	1.937	.803	.856	.053	.028	-.184	.055
Target:I2a	.057	-.065	-.280	.089	.021	-.503	.206	.283	-.286	-.134	1.130	.301	.024	.803	1.919	.331	.447	1.005	-.075	.403
Target:I2b	-.032	-.246	-.685	-.178	-.015	-.056	-.363	-.084	-.205	.011	-.329	-.203	.186	.856	.331	1.571	.466	-.560	.243	-.698
False:Target:D2	.189	-.011	.320	-.096	-.042	-.254	-.202	-.285	-1.101	-.104	.245	.014	-.740	.053	.447	.466	2.766	.665	.572	-.060
False:Target:I1	.305	.220	-.080	.199	.010	-.541	.020	.325	.068	-.259	.824	.507	-.492	.028	1.005	-.560	.665	2.878	.133	1.345
False:Target:I2a	.459	-.259	.151	-.505	-.272	-.234	-.442	-.635	-.222	.945	.408	.250	.401	-.184	-.075	.243	.572	.133	3.673	1.551
False:Target:I2b	.126	.031	-.282	-.307	-.161	-.365	-.095	-.439	.409	.427	.582	.146	.438	.055	.403	-.698	-.060	1.345	1.551	3.392

Items (with a factor 1,000)

[14] To ease convergence, some models were run with log(RT) divided by a factor 10.

	(Intercept)	False	Target	D2	I1	I2a	I2b	False Target	False D2	False I1	False I2a	False I2b	Target D2	Target I1	Target I2a	Target I2b	False Target D2	False Target I1	False Target I2a	False Target I2b
(Intercept)	.098	-.013	-.018	-.006	-.033	-.035	-.114	-.002	-.078	.173	.059	.119	-.108	.048	-.019	.003	.032	-.276	-.005	.036
False	-.013	.860	-.492	.036	.746	.274	.202	-.199	-.029	-.442	.149	-.429	-.765	-.288	-.290	-.034	.211	.423	.537	.505
Target	-.018	-.492	.405	-.041	-.519	-.115	-.198	.099	.009	.209	-.407	.215	.369	.073	.192	.027	-.212	-.850	-.582	-.376
D2	-.006	.036	-.041	.018	.049	-.020	.052	-.028	-.010	-.005	.063	-.025	-.013	-.001	-.040	-.030	.044	.180	.092	.036
I1	-.033	.746	-.519	.049	984	22.1	15.8	-6.83	.959	.854	-12.3	-10.3	-.963	-22.2	-6.95	6.99	6.67	12.8	3.71	2.42
I2a	-.035	.274	-.115	-.020	22.1	.672	.395	-.153	.076	-.239	-.391	-.409	-.232	-.582	-.164	.209	.155	.182	.184	.100
I2b	-.114	.202	-.198	.052	15.8	.395	1.58	.048	.189	-.480	.058	-.345	.268	-.415	-.641	-.158	.312	1.63	.489	.121
False:Target	-.002	-.199	.099	-.028	-6.82	-.153	.048	.235	.069	.068	.162	.205	.125	.040	.164	-.054	-.117	-.122	-.205	-.186
False:D2	-.078	-.029	.009	-.010	.959	.076	.189	.069	.123	-.135	.039	-.090	.178	-.042	.052	.033	-.057	.316	.025	-.068
False:I1	.173	-.442	.209	-.005	.854	-.239	-.480	.068	-.135	2.02	.634	.759	-.146	-.126	.075	-.319	-.801	-1.13	-.487	-.183
False:I2a	.059	.149	-.407	.063	-12.3	-.391	.058	.162	.039	.634	2.72	.513	-.075	-.044	.153	-.740	.402	1.27	.406	.481
False:I2b	.119	-.429	.215	-.025	-10.3	-.409	-.345	.205	-.090	.759	.513	1.17	.212	.185	-.033	-.090	-.123	-.050	-.420	-.208
Target:D2	-.108	-.765	.369	-.013	-.963	-.232	.268	.125	.178	-.146	-.075	.212	2.25	1.45	.321	.498	-.238	.889	.047	-.117
Target:I1	.048	-.288	.073	-.001	-22.2	-.582	-.415	.040	-.042	-.126	-.044	.185	1.45	2.90	.342	.401	-.194	-.316	.576	.595
Target:I2a	-.019	-.290	.192	-.040	-6.95	-.164	-.641	.164	.052	.075	.153	-.033	.321	.342	1.35	.389	-.460	-1.04	-.195	-.015
Target:I2b	.003	-.034	.027	-.030	6.99	.209	-.158	-.054	.033	-.319	-.740	-.090	.498	.401	.389	.908	-.365	.219	-.376	.064
False:Target:D2	.032	.211	-.212	.044	6.67	.155	.312	-.117	-.057	-.801	.402	-.123	-.238	-.194	-.460	-.365	1.76	1.24	1.06	.203
False:Target:I1	-.276	.423	-.850	.180	12.8	.182	1.63	-.122	.316	-1.125	1.27	-.050	.889	-.316	-1.04	.219	1.24	8.06	2.03	.701
False:Target:I2a	-.005	.537	-.582	.092	3.72	.184	.489	-.205	.025	-.487	.406	-.420	.047	.576	-.195	-.376	1.06	2.03	4.55	.993
False:Target:I2b	.036	.505	-.376	.036	2.42	.100	.121	-.186	-.068	-.183	.481	-.208	-.117	.595	-.015	.064	.203	.701	.993	1.89

Experiment 2

`Correct~TruthValue*Polarity*Training +(1+TruthValue*Polarity|Subject) +(1+TruthValue*Polarity*Training|Item)`[15]

Subjects

	(Intercept)	False	Negative	False:Negative
(Intercept)	.927	-.041	-.351	.338
False	-.041	.081	-.008	-.058
Negative	-.351	-.008	.155	-.063
False:Negative	.338	-.058	-.063	.334

Items

	(Intercept)	False	Negative	SI-Training	False Negative	False SI-Training	Negative SI-Training	False Negative SI-Training
(Intercept)	.446	-.190	-.167	-.153	.018	.199	.003	-.113
False	-.190	.081	.071	.065	-.008	-.085	-.001	.048
Negative	-.167	.071	.173	.092	-.156	-.138	-.104	.361
SI-Training	-.153	.065	.092	.158	-.099	-.182	-.036	.342
False:Negative	.018	-.008	-.156	-.099	.225	.139	.140	-.534
False:SI-Training	.199	-.085	-.138	-.182	.139	.218	.063	-.433
Negative:SI-Training	.003	-.001	-.104	-.036	.140	.063	.095	-.303
False:Negative:SI-Training	-.113	.048	.361	.342	-.534	-.433	-.303	1.383

`Answer~Training*Polarity+TruthValue:Training +(1+TruthValue*Polarity|Subject) +(1+TruthValue*Polarity*Training|Item)`

Subjects

	(Intercept)	True	False	Negative	True:Negative	False:Negative
(Intercept)	4.812	-4.836	-4.273	-1.097	.886	1.135
True	-4.836	5.765	3.333	.805	-.701	-.917
False	-4.273	3.333	4.862	1.304	-1.019	-1.385
Negative	-1.097	.805	1.304	2.432	-2.164	-2.447
True:Negative	.886	-.701	-1.019	-2.164	2.125	2.229
False:Negative	1.135	-.917	-1.385	-2.447	2.229	2.835

Items

	(Intercept)	True	False	Negative	SI-Training	True Negative	False Negative	True SI-Training	False SI-Training	Negative SI-Training	True Negative SI-Training	False Negative SI-Training
(Intercept)	.408	.011	-.569	-.176	-.516	.016	.182	.208	.492	.659	-.528	-.574
True	.011	.147	-.157	-.091	-.227	.098	.273	.236	.217	.438	-.493	-.589
False	-.569	-.157	.930	.331	.928	-.117	-.517	-.517	-.886	-1.330	1.205	1.357
Negative	-.176	-.091	.331	.146	.378	-.066	-.274	-.270	-.370	-.576	.555	.597
SI-Training	-.516	-.227	.928	.378	1.010	-.165	-.678	-.668	-.978	-1.513	1.434	1.576
True:Negative	.016	.098	-.117	-.066	-.165	.066	.188	.165	.158	.310	-.344	-.405
False:Negative	.182	.273	-.517	-.274	-.678	.188	.642	.600	.666	1.146	-1.196	-1.325
True:SI-Training	.208	.236	-.517	-.270	-.668	.165	.600	.572	.659	1.097	-1.126	-1.217
False:SI-Training	.492	.217	-.886	-.370	-.978	.158	.666	.659	.955	1.464	-1.388	-1.513
Negative:SI-Training	.659	.438	-1.330	-.576	-1.513	.310	1.146	1.097	1.464	2.369	-2.329	-2.595
True:Negative:SI-Training	-.528	-.493	1.205	.555	1.434	-.344	-1.196	-1.126	-1.388	-2.329	2.357	2.639
False:Negative:SI-Training	-.574	-.589	1.357	.597	1.576	-.405	-1.325	-1.217	-1.513	-2.595	2.639	3.069

[15]Note that Training is a between-subject factor, hence it is not included in the random effects for Subjects.

`log(RT)~Polarity*TruthValue*Training +(1+TruthValue*Polarity|Subject) +(1+TruthValue*Polarity*Training|Item)`

Subjects (with a factor 10,000)

	(Intercept)	False	Negative	False:Negative
(Intercept)	8.257	-1.046	-.061	1.743
False	-1.046	.141	.018	-.219
Negative	-.061	.018	.833	-.415
False:Negative	1.743	-.219	-.415	1.958

Items (with a factor 10,000)

	(Intercept)	False	Negative	Training-SI	False Negative	False Training-SI	Negative Training-SI	False Negative Training-SI
(Intercept)	.247	.161	.104	-.231	-.018	-1.794	.206	2.004
False	.161	.105	.068	-.150	-.012	-1.166	.134	1.303
Negative	.104	.068	1.094	.284	4.185	2.083	3.002	-.394
Training-SI	-.231	-.150	.284	1.272	.635	5.168	-.751	-5.422
False:Negative	-.018	-.012	4.185	.635	129.436	41.880	86.859	-16.791
False:Training-SI	-1.794	-1.166	2.083	5.168	41.880	42.039	23.842	-33.199
Negative:Training-SI	.206	.134	3.002	-.751	86.859	23.842	59.937	-6.097
False:Negative:Training-SI	2.004	1.303	-.394	-5.422	-16.791	-33.199	-6.097	35.674

`log(RT)~Polarity*Training +(1+Polarity|Subject) +(1+Polarity*Training|Item)`

Subjects (with a factor 10,000)

	(Intercept)	Negative
(Intercept)	5.799	1.124
Negative	1.124	2.861

Items (with a factor 10,000)

	(Intercept)	Negative	Training-SI	Negative:Training-SI
(Intercept)	0.5546228503	-0.2426435961	-0.4217154322	1.7445566158
Negative	-0.2426435961	0.2889190459	0.2047755093	0.5481843359
Training-SI	-0.4217154322	0.2047755093	0.5440212211	1.4590472249
Negative:Training-SI	1.7445566158	0.5481843359	1.4590472249	76.951530808

`rRT~Polarity*Training +(1|Subject)+(0+Polarity|Subject) +(1+Polarity*Training|Item)`

Subject (with a factor 10,000)

	(Intercept)	Negative	
	(Intercept)	Affirmative	Negative
(Intercept)	749.08		
Affirmative		2.5802052812	-8.0203724208
Negative		-8.0203724208	192.3407096883

Items (with a factor 10,000)

	(Intercept)	Negative	Training-SI	Negative:Training-SI
(Intercept)	22.115	-.410	-6.837	-7.514
Negative	-.410	71.433	-32.269	-3.974
Training-SI	-6.837	-32.269	24.085	7.055
Negative:Training-SI	-7.514	-3.974	7.055	16.446