Eye movement evidence for context-sensitive derivation of scalar inferences

Stephen Politzer-Ahles[1,2]*, E. Matthew Husband[2]

[1]*Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong*

[2]*Faculty of Linguistics, Philology & Phonetics, University of Oxford, United Kingdom*

**Address correspondence to*

Stephen Politzer-Ahles

Department of Chinese and Bilingual Studies

The Hong Kong Polytechnic University

Kowloon

Hong Kong

E-mail: sjpolit@polyu.edu.hk

Eye movement evidence for context-sensitive derivation of scalar inferences

A scalar expression like *some* can optionally have an enriched interpretation (approximately meaning "some, but not all") depending on the context in which it appears. Numerous experiments using the self-paced reading method have found evidence that context has an online effect on the interpretation of a scalar term, resulting in faster or slower reading times for a later phrase whose comprehension is dependent on the interpretation of *some*. The present study used eye movements to isolate the time course of this process. We find evidence that the reading time facilitation observed in previous studies was driven by early reading measures, with little reading time evidence for an immediate inference-based processing cost at the scalar expression itself, consistent with previous studies. Our results suggest that comprehenders can rapidly commit to enriched interpretations online without cost and that these enriched interpretations are then used to guide the processing of upcoming sentence material.

Keywords: language, psycholinguistics, eye-tracking, pragmatics, scalar implicature

1. Introduction

A substantial part of language comprehension is inferring messages that were not explicitly said. One of the most intensely investigated types of such inferences is *scalar inference*, whereby a speaker uttering a weaker or less informative expression is believed to mean that a stronger or more informative alternative is false. For instance, "all of the cookies" is more informative (more specific) than "some of the cookies" (which, logically speaking, means any nonzero number of the cookies, up to and including all of them);

accordingly, the interpretation of *some* is often enriched, such that a person saying "I ate some of the cookies" will often be understood as meaning "I ate some, *but not all*, of the cookies" (Grice, 1975; for recent reviews see Chemla & Singh, 2014; Noveck & Reboul, 2008; Sauerland, 2012; Sauerland & Schumacher, 2016).

This sort of scalar inference is generally believed to be context-sensitive: in certain contexts, it is less likely to arise, or the inference is made more slowly and effortfully. This has frequently been shown in on-line psycholinguistic studies in which the [context-dependent] interpretation of a scalar expression, like *some of the*, modulates the processing of some downstream expression. Consider, for instance, the two vignettes below (based on Politzer-Ahles and Fiorentino, 2013):

1) a. **Context highly supportive of scalar inference:** Yousef asked Fatima whether *all* of the students had passed the test. Fatima said that some of them had. She added that the rest were planning to retake the class.

   b. **Context less supportive of scalar inference:** Yousef asked Fatima whether *any* of the students had passed the test. Fatima said that some of them had. She added that the rest were planning to retake the class.

In (1a), because the context explicitly introduces a question about whether *all* is true, then *some of them* is likely to be interpreted as meaning "not all of them". Subsequently, a person reading this passage will be aware that there are still some students who have not passed the test. When this reader later reads *the rest* they will be able to comprehend it quickly (either because they have expected this expression already, or because they are able to more easily integrate it into a discourse model which already has a salient group of students who have not passed the test). On the other hand, in (1b),

the context raises a scenario in which knowing that *all* is not true does not provide relevant information to answer Yousef's question (since Yousef only wants to know if the number of students coming is greater than zero). Thus, *some of them* is less likely to be interpreted as meaning "not all of them", given that this interpretation does not answer the question under discussion. Accordingly, a reader may be less aware of a salient group of students who have not passed the test, and therefore will not comprehend *the rest* so quickly.

The above is one example of a sort of contextual manipulation that may influence the derivation of scalar inferences, but there are also many others. For example, scalar inferences are less likely to be realized when the speaker has incomplete information (Bergen & Grodner, 2012; Breheny et al., 2013; Goodman, & Stuhlmuller, 2013; among others), given that a pragmatic derivation of a scalar implicature requires an assumption that the speaker is knowledgeable, i.e., that the speaker knows how many students passed the test. Scalar inferences are also less likely (or unable) to be realized in downward entailing contexts (Chierchia et al., 2012; Hartshorne & Snedeker, ms.; Hartshorne et al., 2015; among others), as these contexts make the non-enriched interpretation (i.e., "at least one of, up to and possibly including all of") more informative than the enriched "not all" interpretation. More global aspects of the context, e.g., what types of alternatives are available in the experimental context, also modulate the availability of scalar implicatures (Degen & Tanenhaus, 2015, among others). It is possible that different types of contextual information are employed in qualitatively different ways during the online comprehension of scalar expressions; the present study focuses on one of the cases

mentioned above, the comprehension of scalar expressions in upward vs. downward entailing contexts.

Upward entailing contexts are those in which a proposition about a set entails a proposition about its superset (i.e., if it's true that "A black dog came", then it's necessarily true that "A dog came"); downward entailing contexts are those in which a proposition about a set entails a proposition about its subset (i.e., if it's true that "No dog came", then it's necessarily true that "no black dog came"). The antecedent of a conditional (i.e., an *if* clause) creates a downward entailing context, and it has been frequently observed that a scalar expression is less likely to be enriched (e.g., *some* is less likely to be interpreted as meaning "not all") when it appears in such a context, compared to when it appears in an upward entailing context (Chierchia et al., 2012; Hartshorne & Snedeker, ms.; Hartshorne et al., 2015; among others). For example, *some* should be less likely to be interpreted as "not all" in (2b) than in (2a):

2) a. **Upward entailing:** Some of the students passed the class, and the rest need to retake it.

   b. **Downward entailing:** If some of the students passed the class, then the rest need to retake it.

In (2b), the *if* clause is actually more informative if the meaning of *some* is not enriched. That is to say, "if at least one and possibly all of the students passed the class" is a stronger generalization, covering more possible situations, than "if some but not all of the

students passed the class" because the former case entails the latter case.[1] Therefore, the interpretation of *some* is often not enriched in this situation, since it would lead to a less informative rather than a more informative utterance.[2]

---

[1] This generalization only holds when *if* is interpreted logically. However, in natural language, *if* is often used to mean "only if" (Geis & Zwicky, 1971); for example, if someone says "The drive takes five minutes – if traffic is good", they generally mean that the drive only can take five minutes when traffic is good (presumably they know that even if traffic is good, there are many other factors that could cause a delay—such as car problems—and thus it is not logically true that if traffic is good the drive will definitely take five minutes). If *if* is given this interpretation, the un-enriched interpretation of *some* is not stronger than the enriched interpretation of *some* in this context. Crucially, though, when *if* is given this interpretation, the *if*-clause is not a downward entailing environment (compare: "You win the contest if you find an egg" entails "You win the contest if you find a blue egg", but "You only win the contest if you find an egg" is entailed by "You only win the contest if you find a blue egg"). Therefore, this does not break the generalization that the enriched interpretation of *some* is less informative than the logical interpretation in downward entailing environments. It does raise the question, though, of whether the present experiment (described below) truly tested the interpretation of *some* in downward entailing environments; if participants used this interpretation of *if*, then the experiment would not be testing what we intended. We thank an anonymous reviewer for pointing out this issue.

[2] In this example, by the end of the sentence, real-world knowledge (i.e., the knowledge that a student who needs to retake a class is probably a student who did not pass the class) also helps the reader interpret *some* as meaning "not all"; our primary interest, though, is how this

--- Table 1 about here ---

The contextual manipulations shown in (1) and (2), as well as others mentioned above, all modulate both the availability of the "not all" interpretation of *some*, and the ease of processing a later anaphor like *the rest*, as described above. Variations of this paradigm have been widely used in experimental pragmatics (see Table 1), and almost all of them have found an effect of context on the comprehension of a downstream anaphor after a scalar expression. That is to say, almost all the above studies found that *the rest* is read more quickly after having read *some* in a context that is more supportive of scalar implicatures (e.g., 1a and 2a), compared to after having read *some* in a context that is less supportive (e.g., 1b and 2b). However, there are still open questions about the nature of the processes underlying this effect. The majority of these experiments have used moving-window self-paced reading (Just, Carpenter, & Woolley, 1982), which is a relatively coarse measure of how long it takes a participant to read a given word or phrase. This method provides only one measure of reading time per word or phrase, and is somewhat unnatural, as participants are only shown part of a sentence at a time and must repeatedly press a button to continue reading. By comparison, measuring reading

---

interpretation arises online during incremental sentence comprehension, before this late disambiguating real-world information becomes available.

time by recording eye movements allows for both more natural reading and, importantly, multiple measures of reading time (Rayner & Sereno, 1994). For instance, different cognitive processes may result in faster initial reading times on a word, more time spent re-reading a word before moving on, or more time looking back to re-read a word after moving on; such differences are not detectable with self-paced reading.

For these reasons, using eye movement measures to shed light on the specific locus of the abovementioned reading time slowdowns is a valuable means to better understand the kind of processing that underlies this effect. Currently, there are multiple possible explanations for why reading times might speed up at *the rest* in the context that is more supportive of scalar inferences. While it seems relatively uncontroversial that this effect is attributable to facilitation by the enriched interpretation of *some* (i.e., the effect is a downstream consequence of having realized the scalar inference), it is not yet known exactly how that interpretation is eventually deployed to ultimately result in faster reading times at *the rest*. On the one hand, the effect might be related to prediction of this particular expression or concept (i.e., after interpreting *some* as meaning "not all", the reader expects that the next sentence will explain what the situation is for the remaining referents) or facilitation of lexical access of *the rest*, in which case this difference might mainly influence measures of early reading processes. On the other hand, the effect might be related to the difficulty of integrating *the rest* into the discourse model, or even to revision of the interpretation of *some* as meaning "not all" (i.e., enriching the meaning at this late point in the sentence). In these cases, the difference might mainly influence measures of late processes (although this assumption is not uncontroversial; see

Discussion); in the latter case, it might also result in more eye movements from *the rest* back to *some* as readers reconsider their interpretation of the quantifier.

Thus far, only Lewis (2013) has tested the context-sensitivity of scalar inferences using eye-tracking with this paradigm. In that study, the reading time facilitation at *the rest* was driven by differences in late reading measures: specifically, re-reading time (the sum of the durations of every fixation on *the rest* after it had been passed once) and total time (the sum of the durations of all fixations on *the rest*). However, as that study had somewhat different aims (in addition to measuring this effect, it was also focused on testing other contextual manipulations and other types of linguistic scales), there are still open questions regarding how much of this effect was due to scalar inferences in particular. In the paradigm typified in (1) and also used by Lewis (2013), there are multiple differences between the highly inference-supporting context (1a) and the less inference-supporting context (1b). For instance, reading time differences later in the sentence might be due to some other downstream effects elicited by the different contexts themselves, rather than being due to the interpretation of "some of" per se. In many experiments using this paradigm or variations thereof (Bergen & Grodner, 2012; Hartshorne & Snedeker, ms.; Hartshorne, Liem Azar, Snedeker, & Kim, 2015; Politzer-Ahles & Fiorentino, 2013), control conditions are included to replicate the context difference while removing the scalar inference difference; e.g., (3a-b) has the same context manipulation as (1a-b), but uses the critical quantifier *only some*, the interpretation of which is not dependent on context (unlike *some*). In these experiments, evidence for a context-based effect of scalar inferences takes the form of an interaction,

such that there is a context effect on reading times for *the rest* in *some* passages but not in *only some* passages.

3) a. Yousef asked Fatima whether *all* of the students had passed the test. Fatima said that only some of them had. She added that the rest were planning to retake the class.

   b. Yousef asked Fatima whether *any* of the students had passed the test. Fatima said that only some of them had. She added that the rest were planning to retake the class.

The purpose of such *only some* control conditions is to attempt to rule out the possibility that a context effect (i.e., different reading times for *some* or for *the rest* in *all* vs. *any* contexts, or in upward vs. downward entailing contexts) is due to general effects of the context itself, rather than specifically to the context's effects on scalar implicatures. While this type of control may still not fully eliminate potential confounding differences between the contexts (Barbet & Thierry, 2016), it at least allows for a stronger argument that the observed effects are based on scalar implicatures, compared to the argument that could be made without such a control. Thus, while Lewis (2013) provides useful prior information about which reading time measures we might predict to show the effect of interest, it remains an open question whether the facilitation effect at *the rest* in this particular study can be attributed to scalar inferences. The present study, therefore, aims to examine which reading time measures are modulated by scalar inference processing in the full factorial research design used by the majority of other experiments in this area.

A secondary goal of the present study is to examine whether context influences the reading time for *some* itself. There is a longstanding debate over whether scalar

inferences are realized rapidly and effortlessly, or slowly and with a cognitive cost (see Chemla & Singh, 2014, among others, for review). Under the latter hypothesis, *some of them* is expected to be read more slowly in the context that strongly supports scalar inferences (1a), given that such a context requires readers to realize a more specific interpretation of this expression. Under the former hypothesis, on the other hand, *some of them* is not expected to be read more slowly in this context. Empirical results regarding this question are mixed. Three studies have observed such a pattern (see Table 1), but these results are also subject to alternative explanations. The finding from Breheny and colleagues (2006) were probably due to a repeated name penalty evoked by their stimuli, as has been argued previously (Hartshorne & Snedeker, ms.; Lewis, 2013; Politzer-Ahles & Fiorentino, 2013). Regarding the finding from Bergen and Grodner (2012), Lewis (2013) has suggested that it may be due to having to infer a relevant set, rather than to realizing a scalar inference per se. Many stimuli were of the form "This morning, I took attendance at an important meeting with the manager. Some of the company's accountants were there." Thus, the referent in question was not explicitly identified earlier in the discourse and the process of connecting the referent to the discourse may have been costly. Finally, in Lauter (2013), the scalar inference was made explicit by orthographically stressing the quantifier (i.e., *SOME of them* as opposed to *some of them*); thus, the longer reading times may have been due to orthography rather than due to the cost of making a scalar inference. Overall, even if these three results are taken at face value, the state of the field is still such that evidence is mixed regarding whether or not realizing a scalar inference elicits an immediate processing cost. Therefore, in addition to providing fine-grained detail about the reading time effects at *the rest*, the present study

will also provide additional data regarding whether or not this context manipulation elicits a reading time slowdown at *some* itself, in a design that allows a potentially more direct comparison than these previous studies did.

2. Methods

2a. Participants

Data were collected from 51 native English speakers at the University of Oxford and the Oxford community. Data from three participants who frequently dozed off during the experiment were excluded from analysis, as were data from one participant who reported having mild dyslexia, leaving a total of 47 participants (35 women, 12 men; age 18-55, mean age 24.1) in the analysis. Individual demographic information for the participants is available in Supplementary File 1. All participants provided their informed consent and were paid for their participation. Methods were approved by the Central University Research Ethics Committee at the University of Oxford.

2b. Materials

The present study used a manipulation of entailment polarity (2a-b), following Hartshorne (Hartshorne et al., 2015; Hartshorne & Snedeker, ms.), rather than a manipulation of information structure (1a-b). The reason for this was that this manipulation allows for single-sentence stimuli, rather than multi-sentence stimuli like

those shown in (1, 3); this made it easier to mix these stimuli with single-sentence stimuli from other experiments within the same recording session.

The critical stimuli (listed in Supplementary File 2) were 48 sentences adapted from Hartshorne and Snedeker (ms., Experiment 1), following the template shown below. "^" indicates where the sentence was segmented into regions of interest; this character was not actually shown in the experiment.

a) **Upward-entailing, *some***: Isabella recommended^ some of^ the applicants^ to the hiring director,^ and the rest^ didn't meet her criteria.
b) **Downward-entailing, *some***: If^ Isabella recommended^ some of^ the applicants^ to the hiring director,^ then the rest^ didn't meet her criteria.
c) **Upward-entailing, *only some***: Isabella recommended^ only some of^ the applicants^ to the hiring director, ^ and the rest^ didn't meet her criteria.
d) **Downward-entailing, *only some***: If^ Isabella recommended^ only some of^ the applicants^ to the hiring director, ^ then the rest^ didn't meet her criteria.

As described in the Introduction, *the rest* is predicted to be read more slowly in the downward-entailing than the upward-entailing condition—but mainly in *some* sentences, not *only some* sentences. A reading time slowdown on *the rest* can only be attributed to an enriched interpretation of *some* if it appears only in *some* sentences, or is greater in *some* than in *only some* sentences. On the other hand, if both kinds of sentences show similar reading time slowdowns on *the rest* in the downward entailment condition, then that might be occurring just because the conditional itself causes sustained processing cost over the rest of the sentence, for whatever reason.

Downward-entailing clauses were used to provide a context in which scalar implicatures are less supported. It should be noted that there is disagreement on how scalar implicatures could be derived in such contexts. Under a purely pragmatic account, a "not all" inference in this context cannot be derived via Gricean conversational implicatures (because the sentence with the enriched "not all" reading would be less, rather than more, informative than the sentence without it); it can be realized via other routes, however (Geurts & van Tiel, 2013). A "not all" reading realized in the downward-entailing clause is presumably not an implicature at all, given that it may be derived by other mechanisms. Nonetheless, it seems to be an empirical fact that, for whatever reason, the "not all" reading is less supported or less available in this context than in the upward entailing context, all else being equal. The primary goal of the study is to examine how downstream reading times on *the rest* are affected when the preceding scalar implicature was more or less available (under the assumption that in downward-entailing contexts the scalar implicature is less available or completely unavailable), and thus this manipulation was considered appropriate for that purpose.

The factors Quantifier (*some* vs. *only some*) and Entailment (upward-entailing ["clause 1, *and* clause 2"] vs. downward-entailing ["*If* clause 1, *then* clause 2"]) were factorially manipulated to yield four conditions. *And*/*then* and *the rest* were combined into a single region since the connective was short and frequently skipped (on 51% of trials, in a preliminary analysis of the data from 35 participants, no fixation occurred on

and/*then* the first time it was passed), and because when viewing the connective the reader was likely also able to get a parafoveal preview of the critical region.[3]

There were also 83 filler stimuli, including 48 items with the same structure and manipulations of Quantifier and Entailment as the critical sentences but not including *the rest*, and 35 items using other quantifiers in the place of *some* or *only some*: nine each of *all* and *none* in upward-entailing contexts, nine of *all* in downward-entailing contexts, and eight of *none* in downward-entailing contexts. These fillers served to make sure that *the rest* and *some* or *only some* were not completely predictable in their respective positions, and to establish a contrast between relevant quantifiers in the experimental context. The session also included 48 items from a separate experiment on semantic enrichment and 104 items from an experiment on morphosyntactic prediction; none of these items included *if-then* constructions, *the rest*, or the quantifiers used in this experiment.

2c. Procedure

The experiment was conducted on an Eyelink 1000 system with a chin rest. Before the beginning of the experiment, and during the experiment whenever necessary, the participant completed a nine-point calibration. Viewing was binocular, but only the right eye was tracked (except for one participant, for whom the left eye was tracked because the right eye was not tracked well by the system). Each trial began with a drift

---

[3] This analysis was not planned before data collection. The decision to combine these into a single region was made after having seen the data for the first 35 participants.

correction, during which the participant had to fixate on a dot (located at the left boundary of where the sentence would appear) before the trial proceeded. All sentences fit on a single line on the screen, and were presented in black Courier New text against a light gray background. To finish reading the sentence and reveal the comprehension question, the participant had to fixate a dark gray box in the upper-right corner of the monitor.

Each sentence was followed by a comprehension question, which appeared after the participant fixated the gray box. The question was presented along with two possible choices and the participant made their response using the arrow keys on the keyboard. The experiment began with 6 practice trials to acclimate the participant to the procedure, after which the 283 remaining items (critical items and fillers from this experiment, as well as items from the two other experiments) were presented in a fully random order, divided into eight blocks with self-paced breaks in between. The stimuli were organized into 24 lists according to a Latin Square design, such that each participant saw 48 critical sentences (12 per condition).[4] Overall the experiment session lasted from 50 to 80 minutes, including setup and debriefing.

---

[4] While the present study only had 4 conditions, the other two experiments from which items were presented had 6 and 8 conditions; therefore, 24 was the lowest common multiple. For the purposes of the present experiment, however, many of the lists were identical—for example, lists 1, 5, 9, 13, 17, and 21 all consisted of the same items associated with the same conditions in this experiment, and only differed in terms of the item-condition pairings in other experiments.

2d. Eye movement measures

Data were cleaned in four steps (SR Research, 2014): first, fixations of 80 ms or shorter were merged into a neighboring fixation of greater than 80 ms within 0.5 degrees horizontally (if both the preceding and following fixation were longer than 80 ms, the short fixation was merged to the longer of the two); second, the same process was repeated with a duration threshold of 40 ms and a distance threshold of 1.25 degrees; third, in interest areas that had at least three fixations of 140 ms or shorter and none of longer than 140 ms, the short fixations were merged into one; and last, remaining fixations shorter than 140 ms or longer than 800 ms were deleted. These values were based on the defaults in the Eyelink Data Viewer program.

We analyzed the following eye movement measures, mainly based on Lewis (2013):

- *First pass time* (also known as gaze duration): The sum of all fixations within a region from when the region was first entered until when the region was exited in either direction.
- *Go-past time* (also known as regression path duration): The sum of all fixations (including fixations in previous regions) from when the region was first entered until when the region is exited to the right (i.e., until a fixation at a later region is made).
- *Selective go-past time* (also known as right-bounded time): The sum of all fixations on the region in question until the region is exited to the right. In other words, go-past time without including fixations on previous regions.

- *Re-read time*: The sum of all fixations on the region in question after it has been exited to the right; in other words, total time minus selective go-past time. (Note that in the literature "re-reading time" is also sometimes used to refer to a different measure, the sum of fixations after exiting the region to either direction—i.e., total time minus first-pass time.) We only included regions with nonzero re-read times in this analysis.
- *Total time*: The sum of all fixations on the region in question.
- *Regressions in*: Whether or not the region was re-fixated after being exited to the right. (While this measure and the regressions-out measure are reported as percentages in the results below, they were treated as binomial variables in the statistical analysis.) This measure was not used in Lewis (2013), but we included it here to account for trials that were not re-fixated (given that it is possible, for example, for a given region to take equal amounts of time to be re-read in two conditions, but to be re-read more often in one condition than another). Furthermore, this was also included to test the possibility, mentioned above, that reading *the rest* after a less inference-supporting context triggers participants to realize the scalar inference late and, possibly, look back at *some* more frequently in the process of making this re-interpretation.
- *Regressions out*: Whether or not other regions to the left were re-fixated after this region was viewed. This measure was not used in Lewis (2013), but it is a relatively commonly analyzed measure.

Trials in which the comprehension question was answered incorrectly were excluded from analysis. Regions for which the first fixation in that trial was not progressive (i.e.,

regions that were skipped, such that the first incoming saccade [if any] came from a later region rather than an earlier region) were also excluded from analysis.

2e. Statistical analysis

The factors Quantifier and Entailment were each sum-coded (with values of -0.5 for *only some* and for downward entailment, and 0.5 for *some* and for upward entailment) so that their coefficients would correspond to main effects. This means that faster reading times in the upward entailing context correspond to negative coefficients for Entailment, and if that pattern is larger in *some* than *only some* sentences the Quantifier*Entailment interaction will have a negative coefficient. For linear models, the outcome variables were transformed if necessary (models were calculated with raw, square-root, log, or reflected-reciprocal transformed data, and whichever model had the least skewed residuals was used; the analysis code in Supplementary File 4 shows which transform was ultimately used for which measure)[5] and then z-scored; z-scoring was done so that

---

[5] While this means that different reading time measures were analyzed in different ways, it is unclear to use whether this should be considered an advantage or a disadvantage. On the one hand one might argue that the same models should be used in order to be comparable, but on the other hand one might argue that models calculated on datasets which fit the model assumptions to different degrees are also not comparable. For our case, the most important thing is that the models were chosen without regard to matters of statistical significance or which one fit our hypotheses the best; the only piece of information used to choose the model was the skewness of the residuals. As the data and analysis code are available in Supplementary Files 3 and 4, the data can easily be re-analyzed using different transforms.

the coefficients would be in standardized units, making it possible to compare the effect sizes of the terms from different reading time measures. Our interest in standardized effect sizes was due to the fact that the research question was less about whether a significant effect would appear (given that we already expected a particular pattern in the reading times at the rest, based on the previous literature), but on which measures would show the *largest* effect.

Coefficients were estimated with linear mixed-effects models (Baayen, Davison, & Bates, 2008) with fixed effects of Quantifier, Entailment, and their interaction. By-subject, by-item, and by-list random effects were fitted, including intercepts and slopes for all model terms (Barr, Levy, Scheepers, & Tily, 2013). Analysis code can be seen in Supplementary File 4.

3. Results

The data are available in Supplementary File 3, and the analysis code in Supplementary File 4. Reading measures for each region are shown in Table 2.

Accuracy was high overall (median 95%, range 88-98%, minimum number of correct trials per subject per condition = 8) and we do not analyze it further.

--- Table 2 about here ---

3a. "and/then the rest" *region*

and/then the rest was read more quickly after *some* in an upward-entailing context than after *some* in a downward-entailing context in all reading time measures, as can be

seen in Table 2. The same pattern, however, also held for *only some*, which indicates that this effect is not wholly due to scalar inferences. The focus of the present study (and the reason for including the *only some* control sentences) was to identify which reading measures showed an interaction such that the facilitation was larger in *some* sentences than in *only some* sentences.

--- Table 3 about here ---

Results from the statistical model are shown in Table 3. At this region, the crucial interaction was significant (in a one-tailed test, given that we were only interested in one pattern of interaction) in first pass times and not in other measures; first pass times also had the numerically highest coefficient. This interaction is shown in Figure 1. The presence of such an interaction, where the context effect in *some* sentences is significantly larger than that in *only some* sentences, conceptually replicates the results observed in self-paced reading (Bergen & Grodner, 2016; Hartshorne & Snedeker, ms.; Politzer-Ahles & Fiorentino, 2013) and event-related brain potentials (Hartshorne et al., 2015); the results further suggest that this commonly-observed effect may have been driven by first-pass reading processes. [6]

---

[6] An anonymous reviewer raised the concern that participants reading so many sentences with *some* in a conditional followed by a context that disambiguates to a "not all" reading might behave unnaturally. Thus, we also did an exploratory examination of reading times on only the first trial of this experiment for each participant. When looking at only the first trials, there is a

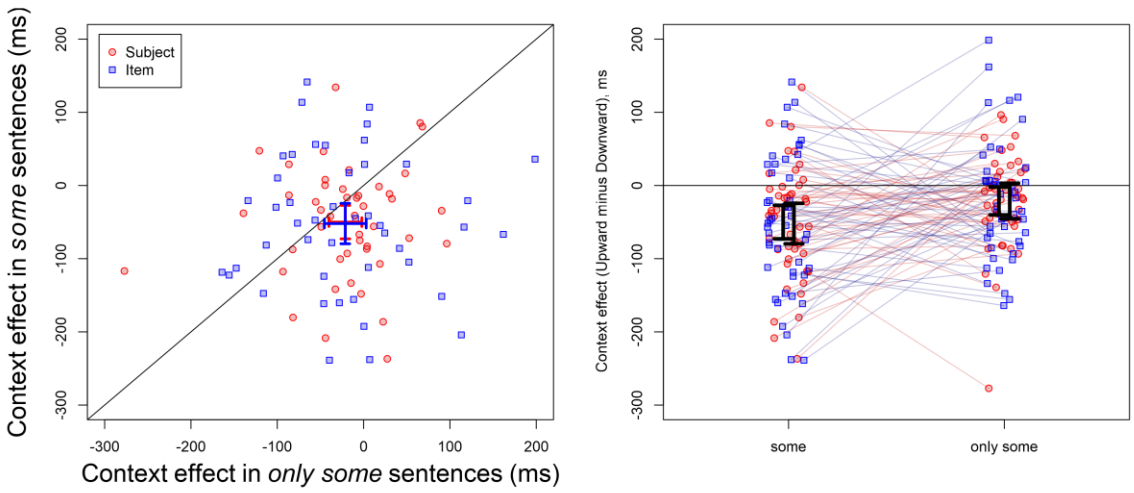

Figure 1. Two visualizations of first-pass reading times at the critical and/then *the rest region. Each point represents either one subject or one item. (Panel A) A two-dimensional visualization based on Rousselet et al. (2016; Figure 1D). The x-axis represents the context effect (reading times in upward-entailing contexts minus reading times in downward-entailing contexts) in* only some *sentences, and the y-axis represents this context effect in* some *sentences. Thus, points in the negative range represent subjects/items for whom* and/then the rest *had faster first-pass reading times in upward-*

---

large difference between first-pass times on *the rest* preceded by *some* in downward entailing contexts (798 ms, SE 141 ms) and those preceded by *some* in upward entailing contexts (688 ms, SE 129 ms), whereas the difference at *the rest* is smaller for sentences with *only some* (downward entailment: 757 ms, SE 110 ms; upward entailment: 729 ms, SE 99 ms). This pattern is numerically consistent with the pattern found in the original analysis (although the overall reading times are substantially slower, as is to be expected early in the experiment before participants have gotten used to the stimuli).

entailing than downward-entailing contexts. Most importantly, points below the diagonal line represent subjects/items for whom this context effect was larger in some *sentences than in* only some *sentences (i.e., the predicted interaction pattern). The red and blue error bars indicate the 95% confidence intervals for the mean of the subject-wise differences and the mean of the item-wise differences, respectively. (Panel B) A more traditional visualization of the same interaction pattern, using connected points to show each subject's or item's context effect in* some *and* only some *sentences. Within each condition, the error bar on the left side is the 95% confidence interval of the subject-wise differences, and the error bar on the right side is the 95% confidence interval of the item-wise differences. (Because this is a repeated-measures comparison, these confidence intervals can be compared against zero to evaluate whether each simple effect is significant, but they cannot be compared against one another to evaluate whether the simple effects are different; Baguley, 2012; Loftus & Masson, 1994).*

As Hartshorne and Snedeker (ms.) propose that the reading time effects on *the rest* in this paradigm may be modulated by the amount of time readers have between the scalar inference trigger *some* and this critical region, we also measured the amount of time between readers' first fixation on the quantifier and their first fixation on the critical *and/then the rest* region. On average this latency was 1813 ms. (For comparison, for the long sentences in Hartshorne & Snedeker, ms., which showed reading time facilitation on *the rest*, the average time was about 2500 ms. For the short sentences, which did not show facilitation on *the rest*, it was about 900 ms.)

3b. Quantifier region

As noted above, a secondary aim of the study was to test whether reading *some* in a context that supports scalar inferences would trigger a processing cost. For this question, we are only interested in measures that correspond to reading times before moving on past the quantifier (first pass time, go-past time, and selective go-past time); later times could be driven by re-reading that happened after *the rest* was encountered, and thus would not be evidence for a processing cost that occurred when the quantifier was first read. We analyzed both the quantifier region and the following region; results from the statistical model are shown in Table 3. At the quantifier, the interaction effect was negligible for go-past times and selective go-past times, and for first pass times it was negligible and in the opposite of the predicted direction (with a larger context effect on *only some* than on *some*). None of these effects was statistically significant. At the region following the quantifier, the interaction effect showed numerical trends in the direction consistent with a processing cost (i.e., slower reading times in upward-entailing than downward-entailing contexts, with this effect larger in *some* than *only some* sentences), although this pattern did not reach statistical significance in any of the three measures.

Thus, the present dataset does not provide strong evidence against the hypothesis that scalar inferences are realized effortlessly. This is consistent with Lewis (2013), Hartshorne & Snedeker (ms.), and Politzer-Ahles & Fiorentino (2013), who also did not find significant reading time slowdowns at the quantifier itself (see, however, Bergen &

Grodner, 2012). Nonetheless, it is worth noting that at the region following the quantifier, all three measures showed numerical trends towards slower reading times for *some* in upward-entailing compared to downward-entailing contexts, with smaller or no trends in that direction for *only some* sentences. On the other hand, in the previous studies that did not find effects at the quantifier, there generally was not even a trend in this direction. Thus, while the data are overall most consistent with the hypothesis that scalar inference does not engender a processing cost at the moment it is realized, they also do not cast serious doubt on the hypothesis that it does (i.e., the study may simply have not had sufficient power to detect such an effect).

To quantify the extent to which the data did or did not support an inference-specific processing cost, we performed a post-hoc analysis using Bayesian mixed models. Unlike frequentist null hypothesis tests, Bayesian models yield a posterior distribution for each parameter, allowing one to make inferences about the likely values of parameters in question. Models, with the same terms as in the analysis above, were fit using the {brms} package in R (Bürkner, in press). The prior for all fixed effects was a normal distribution with a mean of 0 and standard deviation of 0.1, such that most of the prior lay within .3 standardized units of zero. The prior and posterior distributions for the Quantifier*Entailment interaction, for first pass, go-past, and selective go-past times, are shown in Figure 2, along with 95% credible intervals for the interaction coefficient and Bayes factors, which quantify how much the data changed one's confidence in a hypothesis (see, e.g., Wagenmakers, 2007, among others; c.f. van der Linden & Chryst, 2017). As can be seen in the figure, while the posterior distributions for the interaction are all slightly positive, they are not extremely so; only 60% of the distribution is positive

for first-pass time, 67% for go-past time, and 68% for selective go-past time. Given that perfect certainty in the sign of the interaction would correspond to having 100% of the posterior on one side of zero, and perfect uncertainty would correspond to 50%, this suggests that we cannot be very certain about the sign of the interaction (we are closer to being perfectly uncertain than perfectly certain); in other words, the evidence in favor of an interaction effect is not strong. Likewise, the Bayes factors (the ratios of the height of the posterior distribution at a particular point hypothesis, to the height of the prior distribution at that point) for the hypothesis of a zero effect are all close to 1, indicating that the new data has only negligibly changed the confidence in this hypothesis. (Commonly, Bayes factors above 3 (or below 1/3) are taken as indicating that the data have substantially increased (or decreased) confidence in a hypothesis.) Thus, while the eye movements at this spillover region do show a numerical pattern in the direction that would be expected if scalar inferences elicited an immediate processing cost, overall we conclude that there is little evidence that such an effect exists in the population.

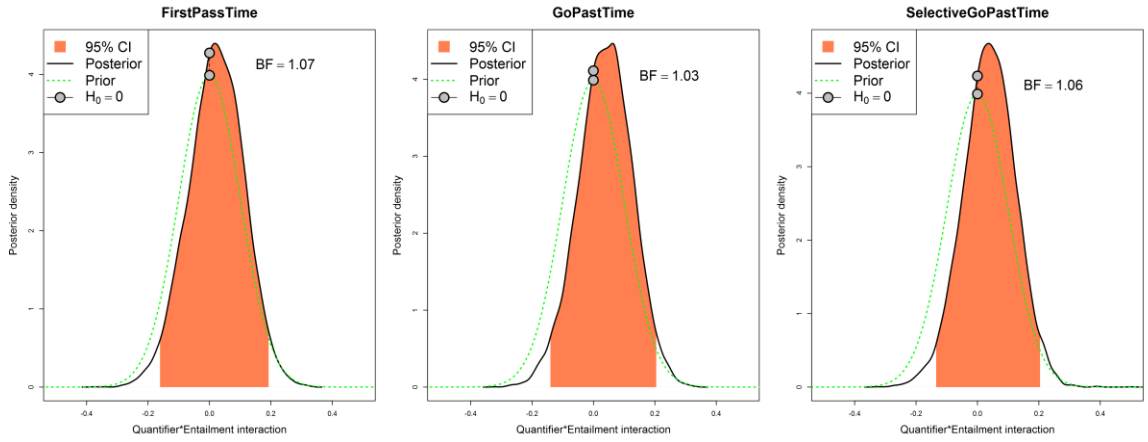

Figure 2. Posterior distributions for the critical interaction effect on first pass times, go-past times, and selective go-past times at the region following the quantifier. Red shaded

regions represent the 95% credible interval of the coefficient. The solid black curve represents the posterior distribution and the dashed green curve the prior distribution; the ratio of these two distributions' densities at 0 is the Bayes factor.

4. Discussion

The present study used eye-tracking while reading to identify the locus of reading time facilitation effects that have commonly been observed downstream of a scalar expression. We replicated the observation of faster reading times for *the rest* after *some* appeared in an upward-entailing context that is more supporting of scalar inferences, compared to when it appeared in a downward-entailing context that is less supporting. Crucially, this pattern was larger in *some* sentences, where the interpretation of *some* is subject to pragmatic context effects, than in *only some* sentences, where the interpretation is semantically fixed. This interaction provides evidence that at least part of the reading time facilitation for *the rest* in the upward entailment condition is due to increased rate of scalar inference realization, rather than just due to declaratives being overall easier to process than conditionals or to other general differences between the upward entailing and downward entailing conditions. This is, to our knowledge, the first study to replicate this pattern of results with eye movement measures. Furthermore, the results suggest that this pattern is due to early eye movement measures (first pass time). Finally, we failed to find strong evidence for a slowdown related to inference-making at the quantifier itself.

The observation that the reading time facilitation is driven by early rather than late reading measures is potentially informative for explanations of the computational locus of this effect. As noted in the introduction, this effect could be explained by early prediction (or facilitation of lexical access) or by late integration, or even by the assumption that encountering *the rest* in the downward-entailing conference triggers an enriched interpretation to be realized late. The present results suggest that the effect is likely to be driven by early processes; it is possible that this is related to prediction, although there may be other candidate explanations as well. Such a pattern would indicate that processing measures at *the rest* in this sort of paradigm do not directly reflect scalar inference-making or meaning enrichment per se, but its downstream consequences (i.e., predictions of upcoming words based on a different interpretation of the scalar expression *some*). This has also been argued to be the case in event-related potential experiments that use brain responses to downstream words to make indirect inferences about the processing of *some*, rather than directly measuring the response to *some* itself (Hunt, Politzer-Ahles, Gibson, Minai, & Fiorentino, 2013; Nieuwland, Ditman, & Kuperberg, 2010; Noveck & Posada, 2003). It should be noted, however, that there was a significant interaction on regressions to the quantifier, such that there were more regressions to *some* in downward-entailing contexts than in upward-entailing contexts (see Tables 2 and 3); this may be consistent with the conjecture that seeing *the rest* causes participants to look back at the quantifier and re-evaluate its interpretation (perhaps by enriching the meaning with "not all"). However, further study (or re-analysis of the present dataset) is necessary to confirm whether these additional regressions to the

quantifier are triggered by seeing *the rest*, as opposed to coming from other parts of the sentence.

A limitation of this conclusion is that the link between various eye movement measures and various cognitive processes is not completely clear (Clifton et al., 2007; Boland, 2004), especially for a topic like conversational implicatures, which has received substantially less attention in eye-tracking research than topics like lexical access and syntactic ambiguity resolution. Therefore, our assumption that early processing measures like first pass time are likely to reflect processes like prediction, and that late measures are more likely to reflect integrative processes, is not uncontroversial. There is some evidence that some discourse-level processing may affect only late reading measures and not early reading measures (e.g., Boland & Blodgett, 2001), or that they may affect both late (including spillover) and early measures whereas more lexical processing may be mostly limited to early measures (Clifton et al., 2007; Staub, 2015). There is substantial variability, however, in which measures are implicated across various studies, and many studies operationalize prediction, integration, lexical processing, discourse processing, etc., in different ways. There is also still general debate regarding how quickly various processes occur in comprehension, not just in the eye movement literature but also in many other psycholinguistic methods. Thus, while the present study provides evidence that the facilitation from scalar inferences on the comprehension of *the rest* happens in early reading measures and presumably early cognitive processes, it is difficult to say precisely which cognitive processes these are.

The failure to find a significant processing cost at the quantifier itself adds an additional piece of evidence to the currently equivocal literature regarding this question.

While many reviews assume that there is convincing evidence that scalar inferences are delayed and/or elicit processing costs (e.g., Chemla & Singh, 2014, among others), the vast majority of studies supporting this claim are those based on end-of-sentence judgments. Many of these are unable to distinguish whether the observed processing costs are directly due to the process of realizing a scalar inference itself, or to subsequent processes (such as ambiguity resolution, or evaluating the inference-derived interpretation of the sentence relative to the context); see, e.g., Bott and colleagues (2012), Chemla and Singh (2014), and Politzer-Ahles and Fiorentino (2013) for review of these issues. When it comes to studies measuring processing costs on-line at the moment the quantifier is read or heard, the results are fairly equivocal. As noted in the introduction, three of the seven extant experiments using this paradigm have observed processing costs at the quantifier, although there may be alternative explanations for each of these effects. It is also possible that the presence or absence of processing cost is moderated by other experimental factors—for instance, there may be reasons why Bergen and Grodner's (2012) manipulation of speaker's epistemic state would elicit measurable processing costs that were not observed in the other experiments manipulating information structure or entailment polarity. In addition to these, Politzer-Ahles and Gwilliams (2015), using magnetoencephalography to measure neural responses in a very similar paradigm as this, found the opposite pattern of processing cost at the quantifier, with greater neural activity elicited in the context that is *less* supportive of scalar inferences. They argued that, rather than an across-the-board cost for making scalar inferences, which is either wholly present or wholly absent, there may rather be a gradient processing cost that can be reduced as a function of context. Studies using other

paradigms have also provided little evidence for an across-the-board processing cost. Barbet and Thierry (2016), using a single-word oddball paradigm, did not find significant evidence from event-related brain potentials that the inference-based interpretation of *some* was more costly to compute, and an unpublished experiment by Politzer-Ahles (ms.), also using a single-word paradigm, failed to find any robust differences between the processing of *some* in a context that required a scalar inference versus a context that did not. Overall, then, the present dataset joins several previous ones in suggesting a parser in which scalar inferences are context-sensitive but not necessarily costly, which is in line with gradient constraint-based proposals of inference processing (e.g., Degen & Tanenhaus, 2015), among others. Nonetheless, there are several potential alternative explanations for the failure to observe an immediate processing cost at the quantifier. In addition to the possibility that there simply is not such a processing cost, it is also possible that there is a processing cost but eye movements are not sensitive to it (e.g., because the costs do not immediately influence the planning and control of eye movements) or that the costs do not occur immediately at the quantifier but rather unfold gradually over the course of the sentence (but, crucially, before *the rest* is read).

Overall, the present study provides a conceptual replication of the observation that context influences the realization of scalar inferences and subsequent processing of a related downstream expression, extending this paradigm into the eye-tracking method. Furthermore, it sheds light on the locus of this effect by revealing that this processing cost may be driven by early reading processes rather than late reading processes. There are many open questions remaining about this effect, such as whether it generalizes to

other populations of readers and to other types of scalar expressions and context manipulations.

Study	**Manipulation**	**Measure**	**$N_{participant}$**	**N_{item}**	**Significant effect at *the rest*?**	**Significant effect at *some*?**
Bergen & Grodner (2012)	Speaker knowledge	Self-paced reading	42	24	Yes	Yes
Breheny, Katsos, & Williams (2006, Experiment 3)	Information structure	Self-paced reading	31	18	Yes	Yes
Hartshorne & Snedeker (ms., Experiment 1)	Entailment polarity	Self-paced reading	28	80	Yes	No
Hartshorne & Snedeker (ms., Experiment 2)	Entailment polarity	Self-paced reading	46	80	No	No
Hartshorne, Liem Azar, Snedeker, & Kim	Entailment polarity	Event-related brain potentials	66 (two experiments)	60	Yes	No
Lauter (2013)	Orthographic focus	Self-paced reading	30	32	No	Marginal
Lewis (2013)	Information structure	Eye movements	24	24	Yes	No
Politzer-Ahles & Fiorentino (2013)	Information structure	Self-paced reading	28	48	Yes	No

Table 1. Studies using the design described in (1) to test the on-line context-sensitivity of scalar inferences.

		Subject Isabella recommended	**Quantifier** (only) some of the	**Object** applicants	**Post-object** to the hiring director,	***the rest*** and/then the rest	**End** didn't meet her criteria
First pass	***Only some* DE**	505 [465, 538]	371 [344, 392]	269 [247, 281]	571 [536, 606]	309 [285, 320]	583 [535, 626]
	***Only some* UE**	596 [559, 635]	381 [351, 404]	271 [253, 285]	579 [545, 613]	290 [263, 303]	617 [569, 655]
	***Some* DE**	470 [436, 499]	245 [213, 259]	265 [248, 280]	579 [547, 614]	336 [304, 352]	592 [542, 632]
	***Some* UE**	564 [528, 593]	253 [217, 265]	288 [267, 307]	595 [555, 626]	283 [257, 293]	647 [589, 698]
Go-past	***Only some* DE**	543 [504, 575]	488 [453, 526]	346 [313, 377]	702 [655, 757]	380 [343, 434]	3157 [3001, 3410]
	***Only some* UE**	-	493 [450, 538]	347 [316, 377]	722 [672, 775]	343 [302, 395]	2920 [2722, 3133]
	***Some* DE**	504 [468, 535]	300 [257, 330]	337 [308, 370]	712 [662, 765]	418 [367, 474]	3173 [3016, 3425]
	***Some* UE**	-	317 [265, 343]	372 [334, 411]	733 [679, 782]	344 [303, 395]	3100 [2870, 3355]
Selective go-past	***Only some* DE**	526 [490, 554]	431 [400, 456]	290 [268, 304]	625 [589, 665]	318 [292, 330]	799 [732, 853]
	***Only some* UE**	-	437 [404, 463]	285 [267, 302]	641 [607, 680]	293 [266, 305]	778 [707, 823]
	***Some* DE**	493 [458, 525]	257 [222, 273]	287 [268, 304]	648 [614, 685]	345 [310, 362]	849 [779, 902]
	***Some* UE**	-	270 [229, 281]	309 [284, 332]	667 [626, 700]	289 [262, 300]	864 [781, 927]
Re-read	***Only some* DE**	681 [610, 752]	595 [494, 642]	506 [348, 491]	683 [441, 686]	413 [304, 462]	-
	***Only some* UE**	571 [510, 623]	579 [503, 628]	503 [366, 488]	648 [371, 646]	362 [260, 421]	-
	***Some* DE**	635 [576, 695]	414 [344, 460]	509 [408, 527]	722 [496, 749]	418 [338, 479]	-
	***Some* UE**	611 [512, 700]	420 [338, 457]	535 [423, 541]	778 [481, 801]	387 [280, 453]	-
Total	***Only some* DE**	1172 [1113, 1259]	902 [829, 988]	545 [504, 600]	880 [834, 970]	469 [415, 492]	-
	***Only some* UE**	1122 [1074, 1203]	932 [869, 988]	583 [547, 641]	850 [803, 945]	378 [322, 401]	-
	***Some* DE**	1094 [1045, 1172]	582 [520, 640]	628 [592, 687]	950 [906, 1041]	524 [467, 559]	-
	***Some* UE**	1129 [1054, 1242]	567 [492, 609]	687 [639, 748]	1003 [919, 1105]	408 [352, 430]	-
Regress. in (%)	***Only some* DE**	95	79	51	37	36	-
	***Only some* UE**	95	78	67	41	43	-
	***Some* DE**	92	85	59	32	23	-
	***Some* UE**	92	71	70	44	31	-
Regress. out (%)	***Only some* DE**	-	18	15	13	7	>99
	***Only some* UE**	-	13	14	16	7	>99
	***Some* DE**	-	16	17	14	5	>99
	***Some* UE**	-	13	18	13	6	>99

Table 2. Reading measures (in milliseconds for time measures, and in percentages for regression measures) at each region, and difference-adjusted 95% [percentile] mixed-effect-model-based intervals (Politzer-Ahles, 2017). The intervals can be roughly interpreted as follows: for a pair of conditions, if neither condition's interval contains the other condition's mean, the two conditions are likely (but not guaranteed) to be significantly different in a mixed effect model. For the Subject region, go-past time and selective go-past times are not shown in upward-entailing contexts since this region was the first region in the sentence (and thus go-past time and selective go-past times are the same as first pass time, except in cases where the participant made regressions to somewhere on the screen outside the sentence). For the final region, re-read times are not shown because they are not possible (except in cases where the participant fixated outside the sentence), and total times are not shown because they are the same as selective go-past times. Note that total time does not equal the sum of selective go-past time and re-read time, because the mean for re-read time does not include observations with re-read times of 0. For regression measures, intervals are not shown since these are binomial data.

	Quantifier	**Quantifier+1**	***and/then the rest***
First pass	Intercept: b = -0.09, t = -1.36	Intercept: b = -0.02, t = -0.31	Intercept: b = -0.06, t = -0.79
	Quantifier: b = -0.71, t = -10.92**	Quantifier: b = -0.01, t = -0.18	Quantifier: b = 0.01, t = 0.26
	Entailment: b = 0.03, t = 0.54	Entailment: b = 0.05, t = 1.12	Entailment: b = -0.20, t = -3.90**
	Interaction: b = -0.02, t = -0.18	Interaction: b = 0.10, t = 0.59	Interaction: b = -0.19, t = -1.71**
Go-past	Intercept: b = -0.08, t = -1.15	Intercept: b = -0.01, t = -0.14	Intercept: b = -0.01, t = -0.12
	Quantifier: b = -0.75, t = -12.37**	Quantifier: b > -0.01, t = -0.05	Quantifier: b = 0.03, t = 0.56
	Entailment: b = -0.02, t = -0.47	Entailment: b = 0.05, t = 1.04	Entailment: b = -0.25, t = -3.94**
	Interaction: b = 0.02, t = 0.22	Interaction: b = 0.14, t = 0.88	Interaction: b = -0.14, t = -1.04
Selective go-past	Intercept: b = -0.10, t = -1.39	Intercept: b = -0.02, t = -0.21	Intercept: b = -0.06, t = -0.78
	Quantifier: b = -0.85, t = -13.00**	Quantifier: b = 0.01, t = 0.31	Quantifier: b = 0.01, t = 0.26
	Entailment: b > -0.01, t = -0.02	Entailment: b = 0.03, t = 0.75	Entailment: b = -0.23, t = -4.34**
	Interaction: b = 0.06, t = 0.48	Interaction: b = 0.12, t = 1.15	Interaction: b = -0.15, t = -1.36
Re-read	Intercept: b = -0.10, t = -1.45	Intercept: b = -0.17, t = -1.93*	Intercept: b = -0.06, t = -0.98
	Quantifier: b = -0.42, t = -7.11**	Quantifier: b = 0.15, t = 2.43**	Quantifier: b = -0.05, t = -0.39
	Entailment: b = -0.02, t = -0.29	Entailment: b = 0.03, t = 0.48	Entailment: b = -0.11, t = -0.90
	Interaction: b = -0.06, t = -0.54	Interaction: b = 0.07, t = 0.60	Interaction: b = -0.17, t = -0.72
Total	Intercept: b = -0.06, t = -0.82	Intercept: b = 0.01, t = 0.10	Intercept: b = -0.06, t = -0.84
	Quantifier: b = -0.74, t = -13.92**	Quantifier: b = 0.22, t = 4.45**	Quantifier: b = 0.16, t = 3.18**
	Entailment: b = -0.02, t = -0.33	Entailment: b = 0.12, t = 2.44**	Entailment: b = -0.35, t = -5.82
	Interaction: b = -0.20, t = -2.44**	Interaction: b = 0.04, t = 0.41	Interaction: b = -0.16, t = -1.15
Regressions In	Intercept: b = 1.83, z = 9.47**	Intercept: b = 1.03, z = 3.10**	Intercept: b = -0.92, z = -6.21**
	Quantifier: b = -0.91, z = -4.13**	Quantifier: b = 0.79, z = 4.65**	Quantifier: b = 0.43, z = 3.40**
	Entailment: b = -0.14, z = -0.64	Entailment: b = 0.73, z = 3.50**	Entailment: b = -0.79, z = -4.52**
	Interaction: b = -1.01, z = -3.38**	Interaction: b = -0.54, z = -1.59	Interaction: b = 0.11, z = 0.30
Regressions Out	Intercept: b = -1.95., z = -13.59**	Intercept: b = -1.93, z = -13.24**	Intercept: b = -3.32, z = -13.80**
	Quantifier: b = -0.23, z = -1.10	Quantifier: b = -0.07, z = -0.40	Quantifier: b = 0.41, z = 1.49
	Entailment: b = -0.12, z = -0.75	Entailment: b = 0.06, z = 0.33	Entailment: b = -0.50, z = -1.03
	Interaction: b = 0.34, z = 1.03	Interaction: b = 0.06, z = 0.20	Interaction: b = -0.02, z = -0.03

*Table 3. Results from the statistical analysis at three regions of primary interest. **p < .05 (estimated based on* t *distribution [Baayen, 2008:270]); *.05<p≤.10 (estimated based on* t *distribution). Contrasts were sum-coded (for Quantifier,* some *was coded 0.5 and* only some *-0.5; for Entailment, upward entailment was coded 0.5 and downward -0.5), so effects of Quantifier and Entailment correspond to main effects. Negative effects of Entailment indicate that downward-entailing was read more slowly than upward-entailing; negative effects of Quantifier indicate that* only some *was read more slowly than* some*. Parameter estimates (*b*) for reading time measures are in z-scored units; estimates for regressions are in log odds of making a regression. For the interaction test at* the rest*,* p*-values are one-tailed.*

Acknowledgements

This research was supported by a John Fell grant to EH. We thank Joshua Hartshorne for sharing the stimuli from Hartshorne & Snedeker (ms.) and Hartshorne et al. (2015).

References

Baayen, H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods, 44*, 158-175. doi: 10.3758/s13428-011-0123-7

Barbet, C., & Thierry, G. (2016). Some alternatives? Event-related potential investigation of literal and pragmatic interpretations of *some* presented in isolation. *Frontiers in Psychology, 7*, 1479. doi: 10.3389/fpsyg.2016.01479

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language, 68*, 255-278. doi: 10.1016/j.jml.2012.11.001

Bergen, L, & Grodner, D. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1450-1460. doi: 10.1037/a0027850

Boland, J. (2004). Linking eye movements to sentence comprehension in reading and listening. In Carreiras, M. & Clifton, C. (ed.), *The on-line study of sentence comprehension: Eyetracking, ERP, & beyond*. Psychology Press.

Boland, J., & Blodgett, A. (2001). Understanding the constraints on syntactic generation: lexical bias and discourse congruency effects on eye movements. *Journal of Memory and Language, 45*, 391-411. doi: 10.1006/jmla.2000.2778

Breheny, R., Ferguson, H., & Katsos, N. (2013). Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition, 126*, 423-440. doi: 10.1016/j.cognition.2012.11.012

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition, 100*, 434-463. doi: 10.1016/j.cognition.2005.07.003

Bürkner, P. (in press). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*.

Chemla, E., & Singh, R., (2014). Remarks on the experimental turn in the study of scalar implicature. *Language and Linguistics Compass, 8*, 373-386. doi: 10.1111/lnc3.12081

Chierchia, G., Fox, D., & Spector, B. (2012). Scalar implicature as a grammatical phenomenon. In K. von Heusinger, Maienborn, C., & Portner, P. (Eds.), *Semantics: An international handbook of natural language meaning* (Vol. 3, pp. 2297-2331). Berlin: Mouton de Gruyter.

Clifton, C., Staub, A., & Rayner, K. (2007). In R. van Gompel and X. Vamvakoussi (Eds.), *Eye Movements: A Window on Mind and Brain*. Amsterdam: Elsevier Science.

Degen, J., & Tanenhaus, M. (2015). Processsing scalar implicature: a constraint-based approach. *Cognitive Science, 39*, 667-710. doi: 10.1111/cogs.12171

Geis, M. & A. Zwicky (1971). On invited inferences. *Linguistic Inquiry, 2*, 561–566.

Geurts, B., & van Tiel, B. (2013). Embedded scalars. *Semantics & Pragmatics, 6*, 1-37. doi: 10.3765/sp.6.9

Goodman, N., & Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science, 5*, 173-184. doi: 10.1111/tops.12007

Hartshorne, J., Liem Azar, S., Snedeker, J., & Kim, A. (2015). The neural computation of scalar implicature. *Language, Cognition and Neuroscience*, *30*, 620–634. doi: 10.1080/23273798.2014.981195

Hartshorne, J., & Snedeker, J. (manuscript). *The speed of inference: Evidence against rapid use of context in calculation of scalar implicatures*.

Hunt, L., Politzer-Ahles, S., Gibson, L., Minai, U., & Fiorentino, R. (2013). Pragmatic inferences modulate N400 during sentence comprehension: Evidence from picture-sentence verification. *Neuroscience Letters*, *534*, 246–251. doi: 10.1016/j.neulet.2012.11.044

Just, M., Carpenter, P. & Woolley, J. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General, 111*, 228-238.

Lauter, M. (2013). *If SOME folks are wise, are others otherwise? The interaction of context and emphasis in online scalar implicature processing* (BA thesis). Yale University, New Haven.

Lewis, S. (2013). *Pragmatic enrichment in language processing and development* (PhD dissertation). University of Maryland, College Park.

Loftus, G., & Masson, M. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review, 1*, 476-490. doi: 10.3758/BF03210951

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, *63*, 324–346. doi: 10.1016/j.jml.2010.06.005

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

Noveck, I., & Reboul, A. (2008). Experimental pragmatics: A Gricean turn in the study of language. *Trends in Cognitive Sciences*, *12*, 425–431. doi: 10.1016/j.tics.2008.07.009.

Politzer-Ahles, S. (2017). An extension of within-subject confidence intervals to models with crossed random effects. *The Quantitative Methods for Psychology*. doi: 10.20982/tqmp.13.1.p075

Politzer-Ahles, S. (manuscript). MEG experiment on scalar implicature using simple composition paradigm. Unpublished manuscript. https://osf.io/c9taq/

Politzer-Ahles, S., & Fiorentino, R. (2013). The realization of scalar inferences: Context sensitivity without processing cost. *PloS ONE*, *8*(5), e63943. doi: 10.1371/journal.pone.0063943

Politzer-Ahles, S., & Gwilliams, L. (2015). Involvement of prefrontal cortex in scalar implicatures: evidence from magnetoencephalography. *Language, Cognition, and Neuroscience, 30*, 853-866. doi: 10.1080/23273798.2015.1027235

Rayner, K., & Sereno, S. (1994). Eye movements in reading: psycholinguistic studies. In M. Gernsbacher (Ed.), *Handbook of psycholinguistics.* Academic Press.

Rousselet, G., Foxe, J., & Bolam, J. (2016). A few simple steps to improve the description of group results in neuroscience. *European Journal of Neuroscience, 44*, 2647-2651. doi: 10.1111/ejn.13400

Sauerland, U. (2012). The Computation of scalar implicatures: Pragmatic, lexical or grammatical? *Language and Linguistics Compass*, *6*, 36–49.

Sauerland, U., & Schumacher, P. (2016). Pragmatics: theory and experiment growing together. *Linguistische Berichte, 245*, 3-24.

SR Research (2014). Eyelink Data Viewer User's Manual. Version 2.1.1.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: critical review and theoretical interpretation. *Language and Linguistics Compass, 9*, 311-327. doi: 10.1111/lnc3.12151

Wagenmakers, E. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779-804. doi: 10.3758/BF03194105

van der Linden, S., & Chryst, B. (2017). No need for Bayes factors: a fully Bayesian evidence synthesis. *Frontiers in Applied Mathematics and Statistics, 3*, 12. doi: 10.3389/fams.2017.00012