Replication of "The realization of scalar inferences: Context sensitivity without processing cost"

1. Introduction

The current study replicates Politzer-Ahles and Fiorentino's (2013) study on scalar inferences. Online scalar inference has been a topic under investigation in psycholinguistics. Even though previous studies have examined online scalar inferencing, the exact mechanism and dynamics of the process is still under debate. By and large, scholars have been interested in the interpretation of "some" and to how much extent people perceive "some" to be in different contexts. "Some = not all" is one of the most studied interpretations—although semantically "some" can mean "all", in terms of pragmatics, by uttering "Some students came", one usually implies that "Not all students came" as one would say so (providing information to the best quality and effectively cooperate in the conversation) if all students came, according to Gricean Conversation Maxims. On one hand, the pragmatic inference is cancellable by uttering "Some of the students came. As a matter of fact, all of them did"; on the other hand, the semantic meaning — "at least one" — of "some" is not cancellable, and therefore, it will be infelicitous to say "Some of the students came. #As a matter of fact, nobody came."

There are two major accounts for the "some = not all" inference. Default processing argues that making "some = not all" inference is easy, automatic, immediate, and context-independent; whereas context-driven processing argues that the inference requires 1) specific context and 2) extra effort. By manipulating different contexts and analyzing reading times, Politzer-Ahles and Fiorentino (2013) tested the two arguments of context-driven model. This replication adopts the same approach and investigates whether scalar inferences are context-sensitive and require extra processing cost.

In this online reading-task experiment, if context-driven processing holds true, it is predicted that: 1) the reading time of the complement set — "the rest" of "some" in inference-supporting context should be faster, because if "some" is identified as "not all", participants should be able to access its complement set "the rest" relatively easily (i.e. "some = not all" is realized context-dependently); 2) meanwhile, the reading time for "some" itself as well as the adjacent regions in inference-supporting context should be slower because readers need to infer writers' intention; 3) for the control group — "only some" — that semantically restricts the meaning of "some" to

"not all", no reading time differences should be observed regardless of the context. This study is preregistered on OSF (link).

2. The experiment

2.1 Methods

- Participants

All participants report themselves to be native English speakers. 9 out of 23 participants are male and the rest 14 are female. Participants' reported age ranges from 24 to 59, *mean* = 38.7, *SD* = 12.17. All participants are recruited from MTurk, which is one of the crucial differences from the original study. The researcher had no control over participants during the experiment, and the attitude with which participants took the experiment was unknown. Presumably participants had considerably more distractors compared to a laboratory environment.

- Materials

Same vignettes as Politzer-Ahles and Fiorentino's (2013) were used. There are three types of stimuli: 6 practice trials, 48 critical trials, and 144 filler trials (198 in total). The experiment starts with six practice trials. After practice trials, participants see a screen showing "Great! Now let's start". Practice trials, and the rest 192 trials appear in randomized order. A third of all trials have easy comprehension questions merely to check whether participants read stories carefully enough. For critical trials, as aforementioned, the context is either inference-supporting ("some = not all") – denoted as "upper-bound context" and signified by "all", or not necessarily so – denoted as "lower-bound context" and signified by "any". The quantifier is either of our interest – "some", or as control – "only some". Participants randomly read 1 of the 2 × 2 combinations for each critical trial. Vignettes are cut into 11 to 12 segments. The first two segments are full sentences that provide background information with a specific context condition ("all" or "any"), in which the second segment is a quoted question. From Segment 3 onwards is quoted answer, and Segment 4 is always the quantifier ("some" or "only some") for critical trials. Segment 8 is always the complement set – "the rest" – of the quantifier. Here are some example sentences.

a. **Upper-bound** *some*: Paul and Deb were trying to decide which gym to go to./ Paul asked Deb whether **all** of them had discounts for college students./ Deb said that/ *some* of them/ did./ She added that/ the rest/ would be/ at/ full price.

b. **Lower-bound** *some*: Paul and Deb were trying to decide which gym to go to./ Paul asked Deb whether **any** of them had discounts for college students./ Deb said that/ *some* of them/ did./ She added that/ the rest/ would be/ at/ full price.

c. **Upper-bound** *only some*: Paul and Deb were trying to decide which gym to go to./ Paul asked Deb whether **all** of them had discounts for college students./ Deb said that/ *only some* of them/ did./ She added that/ the rest/ would be/ at/ full price.

d. **Lower-bound** *only some*: Paul and Deb were trying to decide which gym to go to./ Paul asked Deb whether **any** of them had discounts for college students./ Deb said that/ *only some* of them/ did./ She added that/ the rest/ would be/ at/ full price.

For filler trials, 48 were similar to critical trials but without "the rest"; 48 used "all of" as a substitution of critical quantifiers and did not include "the rest"; and the last 48 replaced the critical quantifier with other scalar expressions such as "many of", "most of", "several of", "a few of", "none of", and numbers.

- Procedure

Participants provided online consent at the very beginning of the experiment. They are informed the approximate time to finish this study and are asked to be patient. They are required to read at a natural reading speed and proceed through the vignette by pressing the space bar. Different from Politzer-Ahles and Fiorentino (2013)'s study, I did not replace all the characters with dashes. Instead, participants only see one segment a time without knowing the full length of the vignette. Another difference is that instead of 5 breaks, as I was not sure how exactly the five breaks were implemented in the original study, participants can take a break before starting reading any vignette during the experiment. Namely, there is a transitional screen saying "Please press space bar to continue. You can take a break while you are on this page" between each vignette. However, the accumulative break time cannot exceed 20 minutes because the time limit for completing the experiment is set to 70 minutes. The original instruction also indicated that after the experiment, participants will be asked several follow-up questions, which is also omitted in the current study due to lack of information. Instead, after finishing the trials, participants can fill in an optional survey asking for their age, gender, native language, how they liked the HIT, and whether they have any comments or concerns about the study.

2.2 Results

Instead of calculating the accuracy of each condition, I used accuracy rate to filter data, as indicated in the OSF preregistration. I excluded participants whose overall accuracy rate is lower than 90%. As a result, 13 out of 23 participants were excluded from further analysis. I then removed all practice trials and filler trials, as well as critical trials that are answered incorrectly. Reading times of Segment 1 and Segment 2 were also excluded as they are background-providing long sentences. In the next step, I removed all observations below 150ms or greater than the overall mean of a specified segment, as Politzer-Ahles and Fiorentino (2013) did. The last step was to trim by participant – removing observations out of *mean*(subject) ± 3 × *SD*(subject) range. All remaining 4312 data points were log transformed.

Mixed-effect linear regression was conducted to model data. In the crossed model, I added segments (3 ~ 11 or 12), quantifier (some / only some), and boundedness (all / any) as fixed effects, with participant and sentence as random effects. Significance only appeared in the main effect of several segments. For what we are interested in, i.e. the interaction effect between reading time of Segment 8 ("the rest") and quantifier and boundedness, there was no significance (Segment 8: "some": "any", b = 0.02, SE = 0.06, t = 0.361, p = 0.718, 95% CI: -0.10 – 0.72), as illustrated in the overall reading-time plot for each quantifier. Because the crossed model did not generate significant interaction effect, I did not proceed to nested model as the original work did.

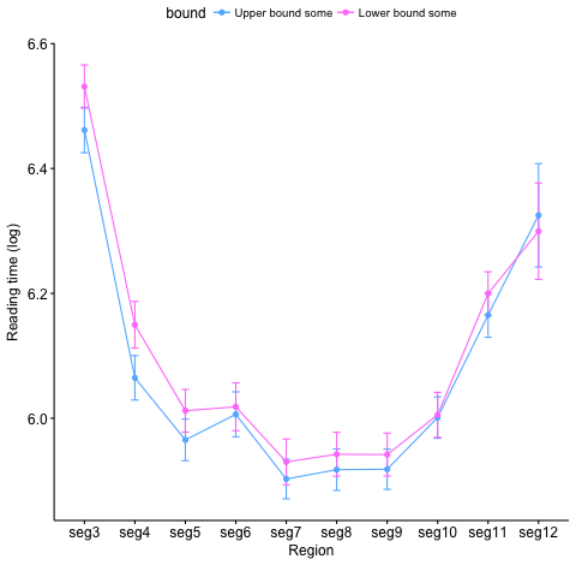

*Figure 1 Segment-by-segment reading times for "**some**"*

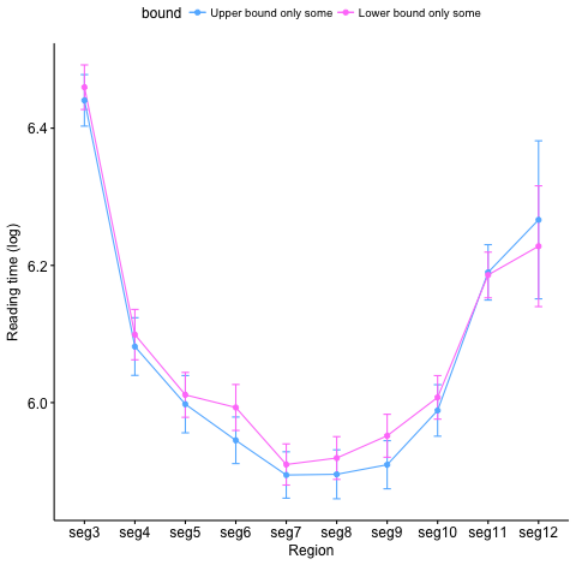

*Figure 2 Segment-by-segment reading times for "**only some**"*

3. Discussion

The current result substantially differs from original study's, which could mainly be due to the difference in experiment condition. Presumably, participants read vignettes much less carefully compared to a laboratory environment where they know that they are monitored in some way. This is reflected in the extremely high rejection ratio—more than half of the participants were rejected due to their poor accuracy rate, and providing that only a third of vignettes come with comprehension questions, it is highly questionable whether they read the rest 2/3 carefully enough. This further suggests that for attention-demanding tasks such as reading vignettes, requiring participants' physical presence at laboratory may be a better choice. In addition, as one participant mentioned in comments, some people may have difficulty keeping track of names while focusing on the content of the story.

An alternative explanation is that the current study in fact rejects context-driven model and is in favor of default processing. In other words, regardless of the boundedness of the context, participants almost always interpret "some" as "not all", which perfectly accounts for the no difference across contexts and quantifiers. As Politzer-Ahles and Fiorentino (2013) argue, if the current study had shown significant context sensitivity (faster reading time of "the rest" in "all" condition and slower reading time in "any" condition) of scalar inference and that default model were used to explain the process, I would have to account for the "inference-cancelling" mechanism demonstrated by the slowdown in reading "the rest". Nonetheless, the current result proves context-independent reading, which is in accord with the definition of default processing.

One of the crucial linking assumptions in this study is that the effort of processing scalar inference should be sensitive to reading time and can be detected by analyzing differences in reading time. Nevertheless, as discussed in the original paper, Bergen and Grodner (2012)'s study provides evidence that reading time can reliably reflect processing effort. Thus, providing the diverging result of the current study, more researches are needed to explore the two models of scalar-inference processing.

Similar to Politzer-Ahles and Fiorentino (2013)'s study, I also investigated the influence of lag time on reading time of Segment 8 ("the rest") and the following segment, as other studies (unpublished data) discovered an interaction effect of context and the lag time between the quantifier and "the rest". The result is same as Politzer-Ahles and Fiorentino (2013)'s: no interaction effect is detected between context and lag time (lag time: lower bound, b = -0.02, SE = 0.04, t = -0.451, p = 0.652, 95% CI: -0.11 – 0.65). Figure 3 corroborates this observation.

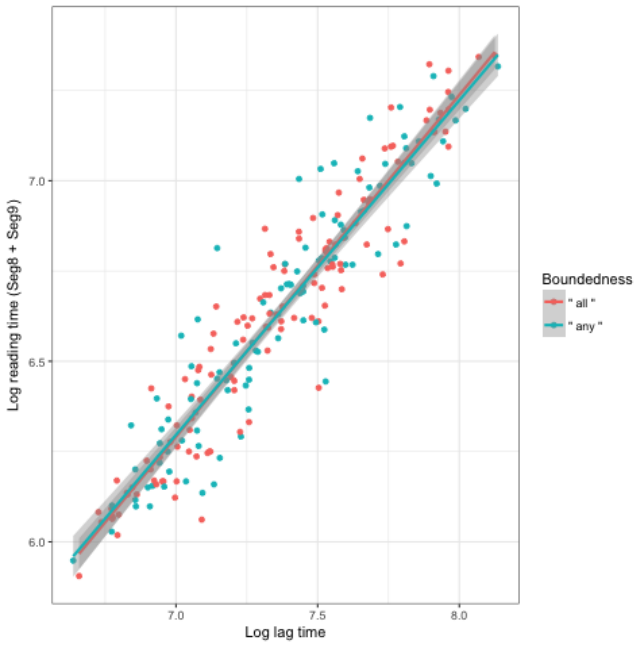

Figure 3 Reading time vs. lag time

4. Conclusion

In conclusion, the present study replicates Politzer-Ahles and Fiorentino (2013)'s work and investigates the context-driven model of scalar-inference realization by analyzing reading times of each segment across different contexts. In general, the non-significant reading-time differences across quantifiers and contexts support neither context sensitivity nor processing cost indicated in the model. The current study raises methodological concerns of using online data-collection platforms such as MTurk in implementing reading-oriented psycholinguistic tasks. On the other hand, the results can be treated as supporting default processing, which complicates the current picture and asks for further studies to test the two models.

5. References

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1450.

Politzer-Ahles, S., & Fiorentino, R. (2013). The realization of scalar inferences: Context sensitivity without processing cost. *PLoS ONE, 8*(5), e63943.