Study information

Title: Ignorance in context: The interaction of modified numerals and QUDs - A replication study

Authors: Mae Grenz, Merlin Marinova, Esma Sakalli, Johanna Venzke

Design for an online experiment concerning ignorance inferences

Background

For centuries, people have been trying to understand each other. Humans are communicators. But it is one thing to speak the language your peers are speaking and it is a whole other thing to extract important implications of what they are actually saying. For instance, when somebody tells you that they saw at least ten rainbows today, you would then from this literal meaning infer that they do not know how many rainbows they saw exactly. This is called an ignorance inference and by going more into depth, the paper “Ignorance in context: The interaction of modified numerals and QUDs” by Matthijs Westera and Adrian Brasoveanu (2014) is challenging us to see if there is indeed a difference between superlative (at least) and comparative (at most) modifiers.

Hypotheses

We are here concerned with the pragmatic account of the ignorance inferences that are associated with superlative but not comparative modifiers (at least vs. more than). Since the authors make a clear distinction between QUDs that are sufficiently fixed and insufficiently fixed, we are going to address these particular research hypotheses:

1. Hypothesis: Ignorance inferences for both modifier types are context-sensitive. More specifically, ignorance inferences depend on whether the question under discussion (QUD) requires an exact answer. While the QUD type HOWMANY requires an exact answer, WHAT does not. Therefore, for QUD type WHAT, we expect higher validity judgements, and lower validity judgements for QUD type HOWMANY.

2. Hypothesis: A contrast in ignorance between the two modifier types of comparative (COMP) and superlative (SUP) will only arise if the context (QUD) is insufficiently fixed. In sufficiently fixed QUDs, ignorance inferences do not depend on the modifier type, i.e., there is no contrast between the two modifier types of COMP and SUP.

As one can notice, there were a few changes made to the hypotheses in comparison to the preregistration report. After revising the original paper one last time, our research group noticed that in the preregistration report, we had not specifically expressed what type of QUD required an exact answer and what type did not (Hypothesis 1). Furthermore, there was no indication in the preregistered Hypothesis 1 about what kind of judgements our research group would expect for the different types of QUDs. Since our research group aimed to investigate the validity

judgements to begin with, it would have been only incomplete to not state the expected results in Hypothesis 1.

Another change was made to Hypothesis 2: In the preregistered report, our research group only wrote down that there would be a contrast between the two modifier types in case of an insufficiently fixed context. However, we had not explicitly written down what would be the case for sufficiently fixed QUDs, namely that the modifier types would not differ from one another. Since the truthfulness of the first part of Hypothesis 2 does not necessarily logically implicate the second one, it was of importance to explicitly say that there would be no contrast between the modifiers with sufficiently fixed QUDs.

Materials. We created our own stimuli after the scheme of the authors. Each item consisted of three components: a question that is asked by a judge (QUD), an answer given by the witness (including a type of scalar modifier, which was either a superlative or comparative - MOD) and a conclusion drawn by the judge based on the conversation beforehand (always an ignorance inference). In our experiment, the factor QUD had three levels and MOD two, resulting in a total of six conditions. QUD ranged over {POLAR, WHAT, HOWMANY} (reference level: POLAR). MOD ranged over {COMP, SUP} (reference level: COMP).

Here is an example of how each of the types could look like:

Judge's question type (QUD):

POLAR: Did you find at most ten of the coins in the wallet?
WHAT: What did you find in the wallet?
HOWMANY: How many of the coins did you find in the wallet?

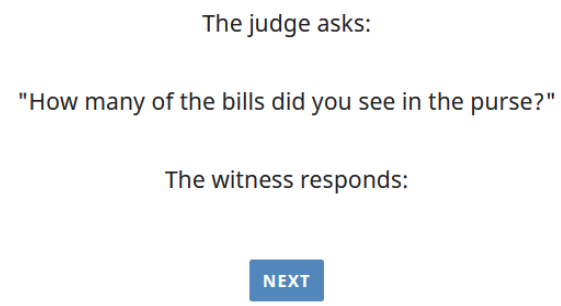
Witness' answer type (MOD):

SUP: I | found | at | most | ten | of | the | coins | in | the | wallet.
COMP: I | found | less | than | ten | of | the | coins | in | the | wallet.

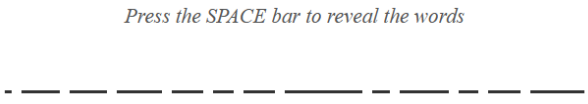
The ignorance inference of the judge:
The witness doesn't know exactly how many of the coins she found in the wallet.

Here is a concrete example of the screens of our study:

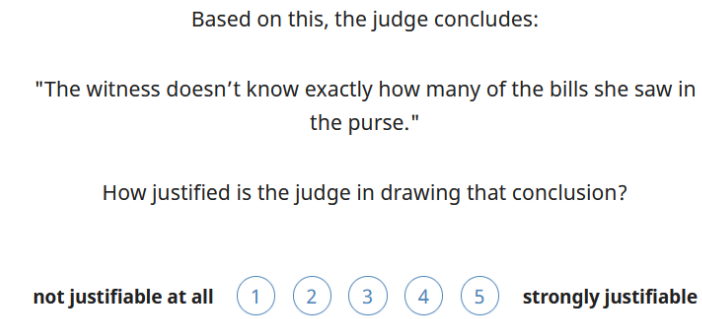
First screen: Judge's question type (QUD):



Second screen: Witness' answer type (MOD):



Third screen: The ignorance inference of the judge:



While creating our stimuli, we also only used the verbs "see, find and hear" like the authors of the original study since we wanted to keep our replication study as similar as possible. The reason as to why these three verbs were used can be explained by the motivation of wanting to portray the witness as an authority over her own perception. The numeral, we chose in all our sentences, was also ten, identical to the original study, with the explanation that ten is a commonly used, round number which is not oddly linked to a specific event (like the superstition of thirteen bringing bad luck). Furthermore, the number ten is a countable number, but still a big enough number to be able to be uncertain about the precise count. Our research group also adapted the courtroom setting in order to keep the situation as natural as possible while asking very similar questions multiple times. This helped decrease the chance of participants worrying about the witness lying or even hiding information. Another

factor we adapted from the original study was how all of the item parts, such as the questions, answers and conclusions, contained prepositional phrases (e.g., “ten of the coins”). This allows us to win time as well as precisibility since effects in tasks like self-paced reading (second screen) are delayed and occur a couple of seconds after the experimental manipulation point. Finally, it is important to mention that we also used only downward-entailing modifier types.

Our research group used 6 items per condition, a total of 36 items with distributions of hear (6), see (15) and find (15). Since the original study had 72 fillers (with distributed verbs hear (12), see (30), and find (30), we also had the exact same proportions. The fillers are there in order to check if the participant is paying attention. The core of the fillers was identical to the items, however, there were a few changes that we made. The changes were as following: In most fillers, in the judge’s question or witness’s answer, the words “approximately”, “probably”, “certainly” and in most fillers’ numeral modifier, the words “only” or “nearly” were added. The fillers also had an extra condition: DID (for example: Did you only see ten bread slices on the kitchen counter?). Finally, the inference of the judge in fillers was also different: It was either “The witness considers it possible that she saw {9,10,11} of the diamonds under the bed” or “The witness thinks the number of diamonds she saw under the bed is comparably high.”

Procedure. The experiment consists of four parts:

(i) introduction

(ii) instructions

(iii) main test phase

(iv) post-experiment questionnaire

We conducted exactly one experiment, which included participants reading a conversation taking place in a courtroom between a judge and a witness.
The second screen, containing the witness’s answer, involved a self-paced reading task where participants were asked to read the answer word-by-word. Here, the SPACE bar revealed the next word while also hiding the preceding one. The reading time was recorded for each word. The third screen was used to measure how justifiable the participants think the judge’s conclusion is.

The order of the procedure was as follows: 6 items of one condition, 12 fillers, 6 items of another condition, 12 fillers and so on and so forth.

In each experiment, there were 6 lists of items. In each list, the items were rotated through the 6 conditions where the items were balanced across conditions and every item on the list appeared exactly only once.

Sample. In total, we had 10 participants, of which 5 were female. We did not need to exclude someone since everybody passed the test by referring to the obviously invalid fillers accordingly. The participants were acquired via round mail to the Cognitive Science community server and through personal contacts. The participants did not receive any VP hours or money for participating.

Implementation. To create the 6 lists, a double Latin square design was used. One Latin Square randomized the order of the conditions across the lists. Another Latin Square randomized the items within one list, satisfying the criteria mentioned in the Materials section. The 6 lists were hardcoded and the participant was assigned to them randomly by the sample function of lodash.

For each participant, the 72 fillers were randomized using a version of the Fisher-Yates-shuffle provided by the shuffle function of the lodash library.

Finally, the item and filler trials were mixed following a block design. 6 different items of one condition were followed by 12 filler trials, leading to a total number of 108 trials per participant.

The experiment was realized with _magpie.