Reproducibility Project: Template for Replication Report
Open Science Collaboration

Replication reports should all use this template to standardize reporting across projects. These reports will be public supplementary materials that accompany the summary report(s) of the aggregate results.

Other useful documents

- Research guide for conducting replication projects
- Executive summary: Detailed description of the reproducibility project
- Possible interpretations of a failure to replicate
- Spreadsheet for documenting replication projects
- Open Science Framework discussion group
- Analysis plan
- Report draft

Replication of “Rational integration of noisy evidence and prior semantic expectations in sentence interpretation” by Gibson, Bergen, and Piantadosi (2013, *Proceedings of the National Academy of Sciences*)

Nicholas P. Moores
npmoores@stanford.edu

Introduction

[No abstract is needed.] Each replication project will have a straightforward, no frills report of the study and results. These reports will be publicly available as supplementary material for the aggregate report(s) of the project as a whole. Also, to maximize project integrity, the intro and methods will be written and critiqued in advance of data collection. Introductions can be just 1-2 paragraphs clarifying the main idea of the original study, the target finding for replication, and any other essential information. It will NOT have a literature review -- that is in the original publication. You can write both the introduction and the methods in past tense.

Methods

Power Analysis

Due to a large original effect size, I elected to reproduce the original sample size of a subset of the experiments, specifically the first three sub-experiments of experiment 1.

Planned Sample

From the original study:

A random order of each experimental list was presented to the participants on Amazon.com's Mechanical Turk, a marketplace interface that can be used for collecting behavioral data over the internet. To constrain the population to American English speakers, we restricted the IP addresses to those in the United States. Furthermore, we asked participants what their native language was, and where they were from originally. Payment was not contingent on answers to these questions. There were 60 initial participants in each experiment, a different set of participants for each experiment (300 participants for experiment 1; 300 participants for experiment 2; 60 participants for experiment 3).

We analyzed only participants who self-identified as native speakers of English from the United States. Furthermore, we only analyzed data from participants who answered at least 75% of the plausible materials correctly.

The sampling method follows the methodology of the original study, with the exception that only the first three sub-experiments of Experiment 1 will be replicated, and thus there will only be 180 participants (60 participants in each sub-experiment).

Materials

From the original experiment:

Twenty sets of materials were constructed for each alternation 1–5 in a 2 × 2 design, crossing construction (alternative 1, alternative 2) with the plausibility of the target alternation relative to the other (plausible, implausible). The items were counterbalanced so that one-half had questions like 6a, in which a "yes" answer indicated the use of literal syntax in interpretation, and the other half had questions like 6b, in which a "no" answer indicated the use of literal syntax in interpretation.

Each set of 20 items was divided into four lists according to a Latin square design, and each list was then combined with 60 filler sentences (e.g., "The commissioner wrote a report for the chairman") to form a presentation list. The target materials were the same across the three experiments. In experiments 1-1 through 1-5, the filler items were all plausible and grammatical sentences. In experiments 2-1 through 2-5, the filler items consisted of the filler items from experiments 1-1 through 1-5, but with 30 of these edited to contain syntactic errors: in 10 items, a function word was deleted (e.g., "The commissioner wrote a report for the chairman." → "The commissioner wrote a report the chairman.); in 10 items a function word was inserted (e.g., "The colonel was knighted by the queen because of his loyalty." → "The colonel was knighted for by the queen because of his loyalty."); and in 10 items, a few adjacent words were scrambled (e.g., *“A bystander was rescued by the fireman in the nick of time.” → “A bystander was the fireman by rescued in the nick of time.”)*

The materials from the original experiment were followed precisely, with the exception that, since only the first three sub-experiments of Experiment 1 are being replicated, there will be no syntactic errors or implausible fillers in the materials, and the materials will not differ by PO-goal/DO-goal or PO-ben/DO-ben.

Procedure

From the original experiment:

Experimental participants were presented with a questionnaire consisting of 60 sentences, like examples 1–5 in Table 1, each followed by a comprehension question, as in 6:

6. a. Active/passive example: The diamond lost the woman.
Did the diamond lose something/someone? (literal syntax: yes).
b. Active/passive example: The ball kicked the girl.
Did the girl kick something/someone? (literal syntax: no).
c. DO/PO-goal example: The girl tossed the apple the boy.
Did the apple receive something/someone? (literal syntax: yes).
d. DO/PO-goal example: The mother gave the candle the daughter. Did the daughter receive something/someone? (literal syntax: no).

The target sentences and the questions were presented simultaneously, and participants could read the sentences and questions as many times as they liked before making their choices. Hence there was no memory component to answering the comprehension questions. (Consequently, the methodology does not distinguish on-line and post-interpretive processes as the source of the effects.)

A random order of each experimental list was presented to the participants on Amazon.com’s Mechanical Turk, a marketplace interface that can be used for collecting behavioral data over the internet.

The answer to the question following each target sentence indicates whether the participant used syntactic or semantic cues in interpreting the sentence. For example, in 6a and 6c, a “yes” answer indicates that the reader used syntax to interpret the sentence, whereas a “no” indicates that the reader relied on semantics, whereas the reverse holds for 6b and 6d.

The procedures from the original experiment were followed precisely, with the exception that, since only the first three sub-experiments of Experiment 1 are being replicated, there will be no syntactic errors or implausible fillers in the materials, and the materials will not differ by PO-goal/DO-goal or PO-ben/DO-ben.

Analysis Plan

The analysis strategy follows the analysis strategy of the original study, which formalizes communication between a speaker and a listener under a noisy-channel model:

$$P(s_i \mid s_p) \propto P(s_i)\, P(s_i \rightarrow s_p)$$

where “s_p is the sentence perceived by the comprehender and s_i is the sentence intended by the producer. $P(s_i \mid s_p)$ gives the probability assigned by the comprehender to any particular hypothesized s_i, given the observed linguistic input s_p. By Bayes’ rule, this can be rewritten on the right-hand side of the above equation as the prior probability $P(s_i)$ that a producer would wish to communicate s_i, times the likelihood of sp given si, which is often notated as $P(s_i \mid s_p)$. We write this likelihood as $P(s_i \rightarrow s_p)$ to make it clear that the likelihood represents the probability of si being corrupted to sp in the process of communication. The prior $P(s_i)$ represents all of the comprehender’s relevant linguistic and world knowledge, including for instance the base-rate frequen- cies of different grammatical constructions and the plausibility of different meanings. This term biases comprehenders toward a priori plausible utterances — things that are likely to be said. The noise likelihood term $P(s_i \rightarrow s_p)$ encodes the comprehender’s knowledge of how sentences are likely to be corrupted during language transmission — for instance, the fact that smaller changes to a sentence are more likely than larger ones.”

The analysis strategy keeps as close to the original analysis as possible. In each sub-experiment, comprehenders’ probability of interpreting the sentence as literally presented is compared using a mixed-effects logistic regression model, with slopes and intercepts by participant and item. Analyses are carried out to test the predictions of the noisy-channel account, that (from the original paper):

1. “Comprehenders should be more willing to forego the literal interpretation when the semantically plausible interpretation involves positing fewer changes to the signal under the noise model, compared with more changes (comprehenders should prefer sentences s_i such that the likelihood of generating s_p, $P(s_i \rightarrow s_p)$ is high. If string edits are independent, then $P(s_i \rightarrow s_p)$ increases as the differences between s_i and s_p decrease, so that s_i is more likely to be hypothesized to be the meaning if s_p can be created from s_i with fewer string edits. For instance, the deletion of a single word should be more likely under the noise model than the deletion of two words.”
2. “The noise model $P(s_i \rightarrow s_p)$ should not treat all changes equally. In particular, comprehenders should infer nonliteral meanings more readily when the change involves a deletion, compared with an insertion. Thus, semantic cues should have a stronger influence for each of the implausible structures in which a word has been deleted from the plausible alternation than for the implausible structures in which a word has been inserted into the implausible alternation”
3. “Because comprehenders do not know the noise rate – the probability that the noise model will corrupt s_i to a different s_p – in every communicative scenario, they must infer it. Increasing the perceived noise rate should encourage comprehenders to infer a nonliteral but plausible alternative.”

4. “Increasing the base rate of implausible sentences should discourage comprehenders from inferring anything other than the literal meaning of the perceived sentence. For example, imagine you are talking to someone who produced many implausible sentences (e.g., a Wernicke’s aphasic patient, or an individual suffering from psychosis). In such a situation, you would be more likely to assume that a particular implausible sentence was intended, rather than produced because of an error. In this case, $P(s_i)$ would be more evenly distributed between implausible and plausible sentences, making comprehenders less willing to deviate from the literal meaning of the observed s_p.”

Unfortunately, since the sub-experiments are only taken from Experiment 1, only the first prediction can be tested, that s_i is more likely to be hypothesized to be the meaning if s_p can be created from s_i with fewer string edits. Thus, if their account holds, people should rely on the literally-presented sentence (rely on literal syntax) more in the major-change alternations tested here (active/passive and locative inversion) than in the minor-change alternation (transitive/intransitive). That is, the percentages in the active/passive and locative inversion experiments should all be higher than all the percentages in the locative inversion experiment.

Differences from Original Study

The differences from the original study all stem from the fact that only the first three-subexperiments of Experiment 1 are replicated. Accordingly, the only syntactic alternations that are being examined are Passive/Active, Obj-Loc/Subj-Loc, and Intrans/Trans, with 60 participants for each sub-experiment for a total of 180 participants. Since only the first three-subexperiments of Experiment 1 are replicated, the fillers contain no syntactic errors or implausible materials. Again, since the sub-experiments are only taken from Experiment 1, only the first two predictions from the original study can be tested, that s_i is more likely to be hypothesized to be the meaning if s_p can be created from s_i with fewer string edits, and that comprehenders should infer nonliteral meanings more readily when the change involves a deletion, compared with an insertion.

These differences are not anticipated to make a difference based on claims in the original article or subsequent published research on the conditions for obtaining the effect seen in the original paper.

(Post Data Collection) Methods Addendum

Actual Sample

sample size = 180 (60 in each sub-experiment), all from the United States. Participants were excluded based on rules spelled out in analysis plan

Differences from pre-data collection methods plan

No differences occurred from the pre-data collection methods plan and the actual data collection carried out.

Results

Data preparation

Data were prepared per the analysis plan. Only data from unique turkers from the United States was analyzed, which resulted in the exclusion of one turker from the sub-experiment on the locative inversion alternation ($N_{locative}$ = 59).

Confirmatory analysis

The analyses as specified in the analysis plan; each sub-experiment was analyzed separately, measuring the proportion of trials during which participants relied on literal syntax when answering the comprehension questions. The resulting proportion reliance on literal syntax by each syntactic alternation type (for the implausible sentences) was calculated and compared to the Gibson et al. data. In addition, in accordance with the original Gibson, et al. analysis, a linear mixed effects model was fit to the data that was as maximal as possible with respect to random-effects structure while still converging.

The final mixed effects model used was:

```
syntax.mixed.model <- glmer(reliedOnSyntax ~ alternation +
plausibility + valence + (1|subjectID) + (1|utterance),
data=ds_model, family=binomial, na.action=na.omit)
```

Exploratory analyses

None

Discussion

Summary of Replication Attempt

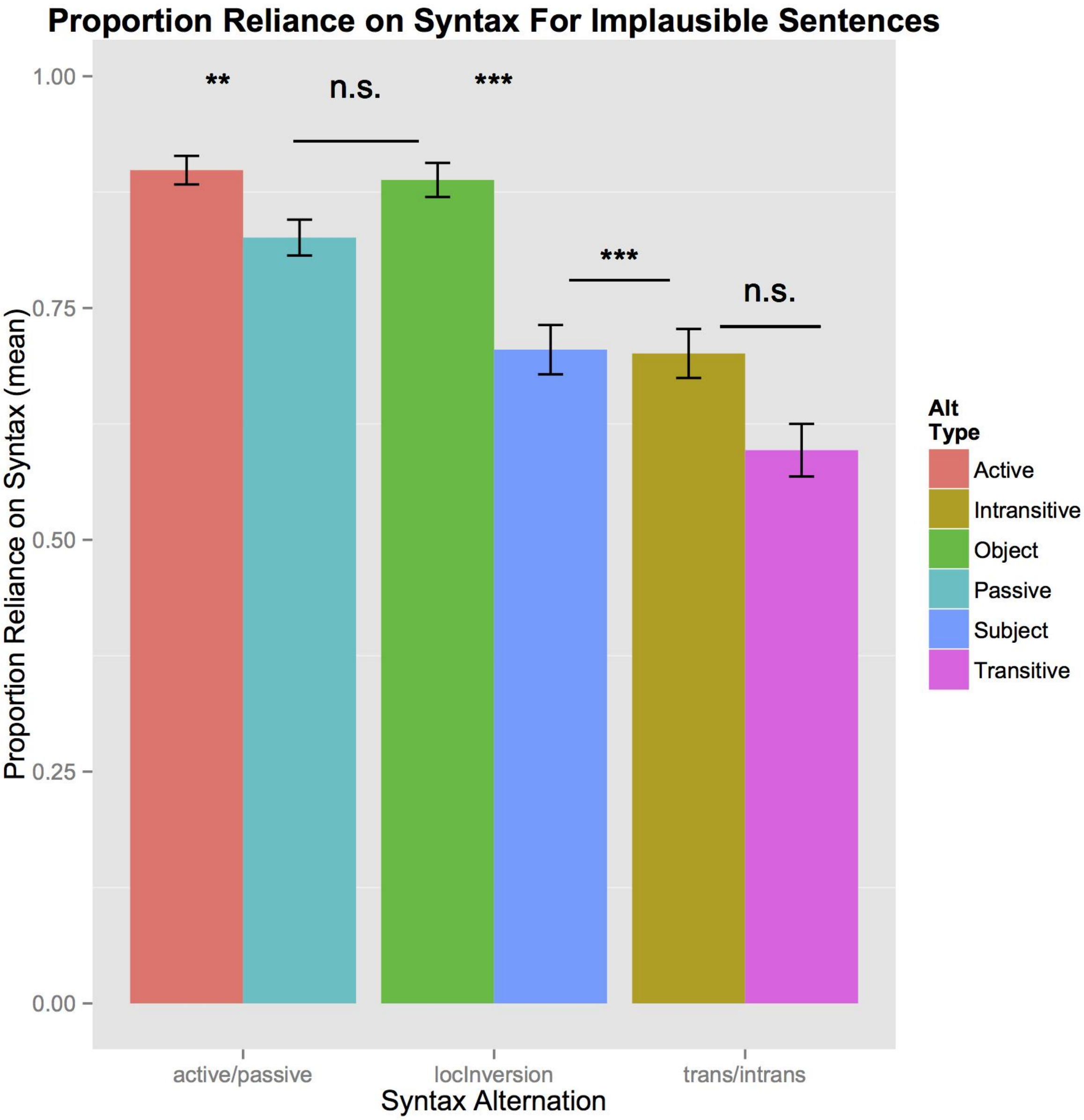
Open the discussion section with a paragraph summarizing the primary result from the confirmatory analysis and the assessment of whether it replicated, partially replicated, or failed to replicate the original result.

The primary result from the confirmatory analysis bore out the result from the original experiment, that participants were significantly more reliant on literal syntax for their interpretation of the sentences when the alternative involved positing more string edits (a major change, as used across the active/passive alternation experiment and the locative inversion alternation experiment), than when the alternative involved positing fewer string edits (a minor change, as in the transitive/intransitive alternation experiment):

- In the original study, reliance on literal syntax by difference in string edits:
  - Major = 93.4%
  - Minor = 56.1% (beta = 3.37, p < 0.0001)
- In the replication, reliance on literal syntax by difference in string edits:
  - Major = 86%
  - Minor = 65% (beta = -1.827, p < 0.0001)

Commentary

One difference between the original study and the replication here is that the linear mixed-effects model used here (which was semi-maximal, but the most maximal model to converge), saw that participants relied on literal syntax significantly more for the active valence of the active/passive alternation compared to the passive valence (beta = -0.78, SE=0.26, z=-2.922, p = 0.00348), and relied on the literal syntax significantly more for the object valence of the locative inversion alternation compared to the subject valence (beta=1.19, SE=0.26, z=4.627, p < 0.00001). This does not necessarily bear on their theory of the noisy-channel account of sentence processing, and so they may also have seen significant coefficients for these alternations but chosen not to report on them if they were significant. From their original paper it is not clear from the plots whether active vs. passive and object vs. subject valences are significantly different from one another. I also do not find a significant difference between participants' reliance on literal syntax between the two valences of the transitive/intransitive alternation, and yet based on their graph of the results of 1.3c and 1.3d, they look to certainly be significantly different from one another.


Proportion Reliance on Syntax For Implausible Sentences
1.00
0.75
0.50
0.25
0.00
Proportion Reliance on Syntax (mean)
**
n.s.
***
***
n.s.
active/passive
locInversion
trans/intrans
Syntax Alternation
Alt Type
Active
Intransitive
Object
Passive
Subject
Transitive