From Tal August:

1. Definitions tested are attached here. def\_text is the text of the definition evaluated, model\_type is the generator of that definition, and reference is the actual definition for that term, term\_id refers to the term in terms.json, also attached here.

2. The interface for complexity is here:<http://scicomm.labinthewild.org/definitions/landing/2>, it is the same for the other human evaluations, but just different questions (detailed in the paper).

3. We used Mechanical Turk for the complexity study, and two trained annotators (an author being one of them) for fluency and factuality. Details are included in the paper.

4. Total cost was $140 for the complexity study. The other two did have a cost, as the annotators were either research assistants being paid separately from the study or authors on the paper.

### Follow up from CT, 31/08/2022 (no response)

Hi Tal,

Many thanks for sending us this information. I had thought I had thanked everyone at the time but it seems I had not replied to yourself, apologies.

We are now at the stage of ensuring that we have enough information to enable our analysis for any reproduction attempts. I therefore have a few more questions about your paper.

1. For the complexity / understandability experiment:
   1. You sent us definitions.json which has 300 lines, matching the number given in the paper. You had 233 participants on MTurk, did some of the participants evaluate multiple items, or, did you get 233/300 responses?
      1. If the former: Was there a limit on how many items each participant could do? (i.e., some of them did a second HIT but no more)
      2. If the latter: Do you know which items in definitions.json did not have results obtained for them?
   2. Did you use any recruitment filters on MTurk? (e.g., geographic region, acceptance)
2. For the fluency, relevance, and factuality experiments:
   1. Can you confirm that both participants annotated all 300 items, for all 3 criteria?
   2. What kind of expertise did the two annotators have? You mentioned one was an author, so they will be an expert in NLP. Was the non-author annotator also an NLP expert?
   3. When you say that annotators were trained, was this specific training for this project, or general training in annotation work? (or both?)
   4. Were any written instructions or worked examples given to annotators prior to the annotation work?
   5. Were the annotators both native speakers?

Thank you again for your time in answering our questions, it is greatly appreciated.

Best regards,

Craig