Manuscript title: Linguistic Complexity Use Among Female Graduate Psychology Students
Invitation to Review date: 26.05.2024
Review submission date: 29.05.2024
Review type: original submission / revision / other
Recommendation: Major Revision

I have reviewed the manuscript “Linguistic Complexity Use Among Female Graduate Psychology Students”. The author(s) explored the relationship between usage of jargon and status. Specifically, they tested the hypothesis that women use more jargon than men in science communication (conference posters).

I found the method of testing the hypothesis (i.e., via computer-based text analysis on conference posters and automated gender assignments) very intriguing. I believe that with such a method come many degrees of freedom that allow for different analysis paths. I think that the manuscript in its current form does not do justice to this flexibility and together with disregarding all open science practices that have been implemented in psychological research, I do not recommend publication of the article. I think the manuscript would profit from a more transparent and detailed description of the text-processing, alternative analyses, and untested assumptions. I also think that the studies should explicitly be presented as *exploratory* (e.g., due to study 2 using 36 different tests and a lack of preregistration). Below, I elaborate on a general issue, list specific remarks, and make recommendations for open science practices. I hope that they can help the authors improve the manuscript.

Sincerely,
Lukas Röseler

General Issue

Given what has been happening in psychology in the last decade (e.g., https://doi.org/10.1177/0956797611417632), I find it daring to report a result with p = .049 (p. 9) and then another one with p = .048. I am speechless about this and frankly would have never expected to read such results in a review of a manuscript in JASNH. Although given my lack of knowledge about the used methods of text analysis, I cannot and I do not accuse you of p-hacking, I hope that you understand that – in the light of the replication crisis and the complete lack of open science practices in your studies (see below) – these findings will also raise other researchers' eyebrows. I see two possible solutions to this issue:

1. Confirmatory approach: You could use data from another conference, preprints, or published journal articles (I suggest preprints due to the existence of an OSF API) and conduct a conceptual and preregistered replication of at least Study 1. In the course of this, I strongly suggest that you adhere to modern standards (e.g., the entire analysis script should be preregistered).
2. Exploratory approach: You could drastically increase transparency of study 1 by reporting results from different analysis paths. Ideally, this would constitute a multiverse analysis (https://doi.org/10.1177/1745691616658637).

Remarks in no specific order

3. I could not find the passage where you cite Anderson, Kraus, Galinsky, & Keltner (2012), which is listed as the second reference. Please note that a finding from this paper could not be replicated as part of Many Labs 2 (https://doi.org/10.1177/2515245918810225). I recommend that if some your research is based on the local-ladder effect, you check out the replication finding. I found the replication by copying and pasting your reference list in the WIP FORRT Replication Database Annotator available here: https://forrt-replications.shinyapps.io/fred_annotator/ (paste references, go to report, scroll down).
4. Given that I could not find a reference to the Anderson et al. 2012 paper in the text, I recommend that you check the reference list for more such cases.
5. When you describe that you used LIWC for the analysis, it is not clear to me to what degree this is computer-based (e.g., is an algorithm rating all texts or is a human involved in some phase?). At this point of the manuscript, I would like to see a brief description of what the way is from the raw text to the scores. Also, I would expect there to be a lot of preprocessing involved. Can you provide details about this?
6. Section "Summary of recent research": I was a bit confused here because it is a mix of recent research and your studies. Some break (e.g., a new headline) or something along the lines of "To investigate X and Y, we conducted two studies. The hypothesis of Study 1 was: …". Maybe the tense has also confused me a bit ("juxtaposition *will* mediate", "hypothesis of Study 1 *is*").
7. Quanteda R package (p. 6): I recommend that you cite all software that you used (ideally with version numbers). For r-packages, you can use the function citation("packagename").
8. Why did you use R, SPSS; and G*Power? I recommend that if you can use R, you may want to do everything with it (or at least drop SPSS as it is not open source).
9. One of the results you report is flagged as inconsistent by statcheck.io: The computed p-value for F(1, 1737) = 5.47, p = .048 is *p* = .01946. Can you please check where this inconsistency comes from? I would expect that it the DFs should be 1549 instead of 1737.
10. I recommend that you adhere to the Journal Article Reporting Standards (JARS) and add confidence intervals to all effect sizes and p-values with three digits. There is a free book on

how to compute effect sizes and Cis and you wouldn't need to use SPSS (p. 9) for it: https://matthewbjane.quarto.pub

11. I suggest that you leave out Table 4, which to me looks like a copy of the SPSS output for bivariate correlations. You wrote above "r(1737) = …", so everything that is needed is already reported.
12. I recommend that you add explanations of the test names in Table 5.
13. Regarding the 36 scores: I think it would be informative to see how they are correlated. Also, you may want to discuss alpha inflation when running 36 statistical tests. Also, I was wondering why Bormuth.GP as a CI of the difference of -1 to +2 million.
14. If the DFs were adjusted for FOG (p. 18), why are you reporting the non-adjusted DFs?
15. "Given that the point estimates of ANOVA effect sizes of all the independent samples tests for gender and linguistic complexity scores are 0 or are in close proximity to 0, the effect sizes tell us that a larger sample size is needed for a stronger effect." (p. 20): This is assuming that there is an effect. First, you may want to consider that there is no effect. Second, with such large sample size and in the case of existing effects (Study 1), you may also want to discuss the crud factor argument (e.g., Meehl, P. E. (2012). Why summaries of research on psychological theories are often uninterpretable. In *Improving inquiry in social science* (pp. 13-59). Routledge.).
16. I recommend that you refine the discussion of Study 2 (p. 27): For example, "linguistic complexity … cannot be accounted for by … status" assumes that gender was actually associated with status, which you did not test for the present sample. I would like to see you discuss alternative reasons why the statistical patterns matches the hypothesis (other than the hypothesis simply being true or false) for both studies.
17. There are a few arguments coming to my mind that I would like see you discussing in the general discussion:
    a. Why did you only look at first authors? Isn't it possible that multiple authors wrote the poster text?
    b. Is there a way to control for other variables such as topic of the poster? I imagine that there could be a lot of noise in the data due to different fields having different frequencies of jargon.
    c. Why did you categorize the generize.io output? When I tried it out, it returned percentages. I understand that they are close to 0 and 1 for most names but there are statistical procedures that can cope with this type of distribution (i.e., logistic models).
18. Minor remarks ("nitpicks", I know that some people do not like reviewers to point out these things but I figured that if I see something like this it can save other people's time [e.g., during copy editing]):
    d. "Barry and Crant" on p. 5 should say "Barry & Crant", I think
    e. This may be a personal preference: "gender does not have an effect on readability scores" (p. 28). Something having an effect on something else implies causality. I suggest a more careful phrasing as gender is not something that can be manipulated while keeping other variables constant, as you write yourself in the first paragraph of p. 29 (e.g. I suggest something like "gender was not associated with readability").
    f. I like the approach of using academic texts (in your case posters) and the automatic coding of names' gender and text attributes. I think a follow-up on your studies could easily be done using pre-prints. For example, you can browse, filter, and download (via an API) titles and abstracts here: https://osf.io/search?activeFilters=%5B%7B"propertyVisibleLabel"%3A"Subject"%2C"propertyPathKey"%3A"subject"%2C"label"%3A"Social%20and%20Behavioral%20Sci

ences"%2C"value"%3A"https%3A%2F%2Fapi.osf.io%2Fv2%2Fsubjects%2F584240da54be81056cecac48%2F"%7D%5D&q=&resourceType=Preprint This way, you could also avoid problems such as multiple authors by using only single-author-papers.

g. "this research proposal" (p. 29): I think this qualifies as a report. In my understanding, a proposal is made for research that has only been planned but not executed.
h. DOIs are formatted inconsistently (doi : … vs. doi.org/…) and sometimes missing (Tourish 2020).

Methodological Checks

19. Usually, I evaluate preregistrations, datasets, code, and materials by multiple factors (e.g., FAIRness, specificity, commenting, etc.). In your manuscript, I could not find any of these things. Unless you report transparently (see also https://dx.doi.org/10.2139/ssrn.2160588) whether one of your studies has been preregistered (and if so, list and discuss all deviations from the preregistration), unless you make your raw and processed data and code openly available, well commented, with a code book, and adhering to FAIR criteria, and allow for reproduction of your results, I do not recommend the publication of your (or anybody's) research article. I am aware that these practices are not required by the journal but I have commited myself to requesting these things as part of the PRO initiative (https://www.opennessinitiative.org/the-initiative/).

Disclaimers:

- I have never conducted computer-based text analyses and do not know the used software. I have little experience with linguistics and the psychology of language.
- I could not conduct a reproducibility check due to the code and data not being available.