GENERAL IMPRESSION

- Timely and interesting topic
- Open science practices (preregistration, transparency, not shying away from null effects)
- Impressively advanced statistics (but sometimes overkill for the studied question?)
- Clearly written for most part (although quite a few missing details, especially on Methods)

GENERAL QUESTIONS / COMMENTS

- **Categoricalness vs Dimensionality**
  It feels like the manuscripts sometimes conflates 'categoricalness' with 'dimensionality'. At multiple places it is stated that the data show evidence of 'categorical perception of gender', where I think is really meant 'one-dimensional perception of gender'. See my slides.

- **Representativeness of the participant samples**
  The first two studies used mainly psychology students as participants if I'm not mistaken. This subpopulation might be more aware of gender-related issues than the general population. Isn't this problematic in terms of generalization of your results? If so, perhaps look for more diverse subject samples in Study 3.

- **Choice of tasks**
  (Related to the previous point) The current studies uses experiments in which participants are explicitly asked to judge the gender of presented faces. This might put them into a 'gender-aware' mindset. It could be interesting (and more 'ecologically valid') when using a task in which subjects are not aware that this is a study about gender categorization.

- **What are gender categorization studies?**
  The manuscripts often refer to gender categorization studies when formulating the problem statement (example: "Gender categorization studies almost exclusively measure categorization allowing participants to respond in terms of woman and men only. This does not capture the diversity of gender"). This is clearly a central term in the thesis. However, it is never described what those studies are. It would be useful to briefly introduce this body of work in the paper Introductions, as it helps understand the reader what it's all about and why misgendering is a serious problem.

- **Is the problem that the thesis aims to address limited to gender categorization studies?**
  (Related to previous point) Isn't 'failure to capture gender diversity' also a problem for studies that look at gender differences? Think of studies looking at gender differences in depression, hypertension, cardiovascular disease, stress, inflammatory bowel disease, etc (when searching for 'gender differences in' on Google scholar, you will get almost 150k hits). Elli's findings potentially also have implications for that body of work.

- **Small inconsistency in data presentation between the two papers**
  Study 1 uses figures where Study 2 uses tables and/or vice versa

MANUSCRIPT 1

- **Abstract**
  Problem statement can be made much clearer: *why* is it a problem that gender categorization studies do not capture the diversity of gender? Is it a methodological problem? Does it hamper progress in theory building? Is it an ethical problem? All of the above?

  The results sentence in the Abstract seems contradictory: providing extra options changes the responses, yet it does not change categorical perception?

- **Four motivations: are they all distinct?**
  p. 5 mentions four "reasons researchers should be concerned about measuring gender categorization with binary options only". What is the difference between (1) and (2) (isn't the minority stress what makes the norm harmful?); and what is the difference between (3) and (4) (isn't the reduction in variation due to the skewing?)

- **Method: figure**
  Please include a figure of the stimuli and the experimental task – would make it so much easier to understand the experiment!

- **Method: how did participants respond?**
  Key presses? Mouse clicks? What about the free-text field? Was the picture present during the response stage? Etc. Lots of missing details in Methods – make sure to include all information required for someone to replicate your experiment.

- **Reasoning behind research question 1**
  You write "This could manifest as either a main effect of condition or an interaction between ...". Is this really true? Wouldn't the interaction require a main effect? (Unless categorical perception in the 'least androgynous faces' would be *increased* by just as much as it is *reduced* in the most androgynous faces)

  A few lines later you write "These questions correspond to main effects of ... and an interaction" – however, I don't see how the two main effects and the interaction map to the three questions

- **Results: individual differences**
  Was the effect that is mentioned at the bottom of p. 10 driven by a few individuals or shared by most individuals? I think it would be interesting to also look at individual data. See my slides.

- **Results: skew**
  p. 11: "Subsequently, we tested whether the inclusion of non-binary response options skewed the distribution of categorization of faces as women and men. For this analysis [...] we tested the outcome variable *binary categorization"*. I was happy to see a confirmatory analysis here

(evidence for the null). However, I didn't understand anything of this analysis. What do you mean with skew here? And how would the model comparison answer this question?

- **Study 1 discussion: no difference between binary condition and free-text condition**
  You write "Thus, the written out choices seem to act as reminders to participants". But what if participants were simply lazy? Writing a text is more work than clicking a button – especially when doing many trials, I would prefer to just click buttons. Or perhaps they didn't know what to write or dare to enter free text fearing that they may write something that goes against accepted norms.

  If you do an experiment again with a free text option, then my suggestion would be to have that as the *only* option (meaning that participants explicitly have to write 'male', 'female', 'don't know', 'non-binary' on every trial, so that giving a response 'non-binary' is not more effortful than giving a response 'female')

- **Hypothesized finding**
  I didn't quite get the logic behind this: "If categorical perception occurs, ratings of woman and man should be skewed near facial femininity = 50. In other words, a face with 33.3% facial femininity would be rated as less woman than that. Therefore, we examined the differences between the two conditions at facial femininity = 33.3% and 66.6%".

- **Methodology: details about responses**
  How were the continuous responses given? Using slider bars? Button presses? And were both responses given in the same screen or were they given sequentially? Etc. Lots of missing details

  Why not label the scales as "Degree of femininity" and "Degree of masculinity"? The "No man ... man" and "No woman ... woman" labels seem a bit odd. Or perhaps it's the same thing. Something you could at least give a little bit of thought.

- **Methodology**
  Impressively advanced methods were used. However, very little is written about the motivation for the chosen methods. For example: why use mixed models instead of Chi square tests / contingency tables? Not knowing the motivation, it feels like a bit of an overkill – (why) weren't simpler methods adequate for answering your questions? (For some questions it might actually be sufficient to simply look at the graphs)

MANUSCRIPT 2

- **Strength of main finding**
  In the abstract you write "Using hen in the writing task increased gender categorizations beyond the binary compared to using other pronouns". As you acknowledge in the manuscript, the effect is rather small. But another important aspect that you do not discuss is how long-lasting the effect is – any ideas on this? Wouldn't this be important to test before getting to excited about this finding?

- **Why a writing task**
  Your title is “Exposure to the Swedish neopronoun *hen* facilitaties gender categorization beyond the binary”. However, your experiment involved much more than mere exposure – isn’t it misleading to frame it like this? (Suggestion: “Use of the Swedish ...” or “Interaction with the Swedish ...”)

- **Subject exclusion**
  53 out of 395 subjects were excluded “because they did not follow the instructions”. How was “follow the instructions” measured? Were these criteria preregistered or made ad hoc? Do the conclusions critically depend on this exclusion or are they similar when analysing the entire sample?

- **Is this really a priming task?**
  You present your task as a priming task. However, isn’t it much more than mere priming? You explicitly tell participants that it is *important* to use the pronoun ‘hen’. (Wikipedia definition of priming: “the idea that exposure to one stimulus may influence a response to a subsequent stimulus, without conscious guidance or intention”)

- **Methods**
  A lot of missing details. For example: was this an online or offline study? Subjects were recruited via an only platform, but the remark about debriefing suggests that it was performed on-site.

- **Reporting of results**
  What are the units for the reported numbers on p. 15 (“Estimate = 0.45, CI = -0.08, 0.97 ...”)?

- **Consistently inconclusive**
  Awesome term :)

- BF10 < 1000 -> i guess this should be BF10 > 1000?

Typos Study 1:

- Page header: RSPONSE instead of RESPONSE
- Abstract: Gender categorizationS studies
- p. 3: “bem_measurement_nodate?”
- p. 4: delete “.” in first line
- p. 4: delete space between “)” and “.”
- p. 5: binary out OF a more varying
- p. 5: “..attempted to address..” -> “attempted to measure” would be more accurate?
- P. 5: “two alternatives: INCLUDING a third gender option, etc”
- P. 9: “were compared the null hypothesis” – something wrong here
- P. 11: “evidence were” -> “evidence was”
- P. 11: “when participants categories faces” -> “... categorIZE faces”

Typos Study 2:

- Abstract: facilitateS
- p. 12: BF01 > 1000 -> should be BF10 > 1000?
- Table 1: there seems to be a typo in 3rd column (‘He’) (proportions add up to >1; and .68 is outside the 95% CI)
- p. 16: BF10 < 1000 -> should be BF10 > 1000?