The effect of response options on gender categorization (provisional title)

Elli van Berlekom[1] & Coauthors[1,2]

[1] Stockholm University

[2] Lund University

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Data & scripts are available at osf link

The authors made the following contributions. Elli van Berlekom: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Coauthors: A lot of things, Author order TBD.

Correspondence concerning this article should be addressed to Elli van Berlekom, Albanovägen 12. E-mail: elli.vanberlekom@psychology.su.se

Abstract

I'm using a premade template & leaving some of their guidlines in place to help me.

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

The effect of response options on gender categorization (provisional title)

The experience of transgender and gender diverse (TGD) individuals suggests that sex/gender is a fluid category which can vary along a wide spectrum. In contrast, social categorization and face perception research often treats gender as a binary consisting of women and men (for example Webster et al., 2004). This is problematic because it indirectly delegitimizes TGD individuals' experiences. Additionally, it may restrict participants' answer, similar to how ratings of age along an old/young binary would restrict and distort ratings of age variation (see Westbrook & Saperstein, 2015; Lindqvist et al., 2019). Furthermore, it may distort answers by communicating ideas about gender. In this study, we aimed to investigate how various gender categorization paradigms influence participants' categorizations of faces.

A cursory glance at the literature on gender categorization reveals that the vast majority explicitly or implicitly suggest to participants that gender consists of the categories woman and man only. The most common method to measure gender categorization is a force-choice task, where participants are presented with a face and the choices are "female" and "male" (see for example, Cloutier et al., 2005; Campanella et al., 2001; Webster et al., 2004; Zhao & Bentin, 2008). A slightly different task asks participants to rate the faces on gender as a quality, rather than a category, often with "feminine" and "masculine" as endpoints on a single scale (e.g. D'Ascenzo et al., 2015; others). Overall, despite some variations, this is a literature where gender is frequently is presented as a binary.

Presenting gender as a binary communicates to participants that the researchers do not view non-binary genders as legitimate. For TGD individuals, this may contribute to a wider pattern of cisgenderism, the ideology that discards people's own conception of their gender identity. Researchers may raise the objection that binary response options may be the most suitable for the research question or the planned statistical analyses. This may be

the case, but it should be weighed against the real harm that is being done by these options.

Furthermore, it is worth questioning whether a binary forced choice is ever the most appropriate method to measure gender categorization. This position seems to be premised on the assumption that there is some fundamental basis to gender, a truth which can be distorted. According to this view, binary is the neutral way to measure sex/gender categorization and anything else is the result of agenda-driven or political motivations. If gender is instead viewed as a social construct, which arise as a result of repeated discourse, this suggests that there is no neutral way to measure gender categorization. Rather, there are multiple alternatives which come with their own limitations and restrictions or suggestions.

Indeed, gender can be measured in many different ways, with drastically varying results. For example, Bem (1974) constructed scales to measure femininity and masculinity as separate personality traits. She found that many people had a mixture of feminine and masculine traits. In another example, when Joel and colleagues (2014) asked ostensibly cisgender participants whether they ever experienced shifts in their gender identity, a sizable group had. Lastly, and Westbrook and Sperstein (2015) showed that there are many potential ways participants answer questions about their gender identities, including rating femininity and masculinity on separate dimensions. When offered these separate sliders, participants generally offered a high degree of androgyny. These results, which primarily regard people's self-categorization and not categorization of others, nevertheless suggest that when people are given the options to categorize gender beyond the binary, they frequently use them.

Additionally, gender binaries can be created or enhanced through statistical practices. For example, Hyde and colleagues (2018) concluded that the statistical practice of examining mean differences between women and men exaggerates the difference and

downplay gender similarities. Hester and colleagues (2020), showed both that perceived differences between the faces of men and women were pronounced when only means were examined, and when gender was measured as consisting of a single dimension with femininity and masculinity at opposing ends. These studies show that when experiments are constructed to take diversity of gender into account, the results often reveal a diversity of gender. This primarily suggests that studies which only measure binary gender are unnecessarily and artificially restrictive. **I added this paragraph earlier, but I'm not sure if it's actually relevant**

If binary response options have this problem, it is worth conceptualizing what are some possible alternatives. One easy solution is the inclusion of a third third gender option, such as "other" or "non-binary". This has the benefit of acknowledging the existence of TGD individuals, which is something that many TGD inviduals have expressed they would like to see (Richards, 2021). However, solely adding a third alternative is not enough. It also indirectly implies that TGD people are androgynous, which is not always the case (ref). What TGD individuals and activists have championed is instead a general caution about gender categorization given that TGD people can present in a wide variety of ways, not all of which are androgynous (ref). Therefore, in a categorization task, it would be preferable to have the option of expressing a sense of uncertainty, possibly through an "I don't know" alternative. Although psychometricians discourage the inclusion of "I don't know" responses on the basis that it discourages participants from taking a stance (Kosnick et al., 2010) from a TGD perspective on gender categorization, this is precisely the outcome which is desirable. Lastly, a way to skirt all of these issues is to allow participants complete freedom to categorize however they choose, using open-ended text entry. *to do: include a sentence bridging to the research questions*

The aim of the present study is to present as a proof-of-concept what categorization studies which are senstive to TGD individuals may look like. As such we have two research questions related to the inclusion of additional response options.

Research question 1: Do people use beyond-binary options when they have them?

Research question 2: Two what extent do beyond-binary responses affect the distribution of woman/man responses?

Categorical Perception & Gender Categorization

this whole section is kind of a work in progress

Another question one might consider about response options is the degree to the implications of response impact participants view of gender. As we discussed, when gender is measured as only the categories "woman" and "man" the implication may be that gender/sex consists of two discrete mutually exclusive categories (ref). Conversely, when gender is not presented as a binary, the implication is that gender can be more inclusive.

One way to consider how response options shape the perception of gender is using the concept of categorical perception. Categorical perception is a perceptual effect where people tend to accentuate the differences of continuous stimuli. It has been observed for colors and for sounds. The existence of categorical perception suggests that people have a strong sense that categories exist. Importantly, categorical perception has been observed for gendered faces (Campanella et al., 2001). However, if participants respond to gender categorization with options that are less binary, maybe they will exhibit less categorical perception?

Research question 3: does a binary slider lead to more categorical perception that two separate sliders?

Experiment 1

Method

Participants

Participants ($N = 68$) were speakers recruited through advertising online and on the university campus (M_{age}= 37.67, SD_{age} = 14.56, Range = 20 - 69). All participants were informed that participation was voluntary. In term of gender, the participants were 35 women, 32 men and 1 who did not indicate gender. Written consent was obtained from all participants.

Material

Faces were produced using faces from the London Face Database (deBruine) and the Chicago Face Database (ref) morphed with on Webmorph (ref). For Black, Asian and White faces, the six most feminine faces of women and the six most masculine faces of men were selected, using the codebook provided by the researchers. The faces were matched, so that the most feminine face were morphed with the most masculine face and so on. The morphs were made in 7 steps, from completely feminine to completely masculine. Because there were 18 pairs morphed in 7 steps, the total number of faces was 126.

Measures

Gender binary beliefs (GBB) were measured with an adapted versoin of the Gender Binary Beliefs scale by Tee & Hegarty (2014). The scale measured the extent to which participants endorsed items such as "*placeholder*" and *placeholder*

Beyond-binary responses represented the categories where participants made a response that were not woman or man. This was a dichotomous variable that was calculated from the categorization data by combining the responses of "I don't know" and "non-binary". These beyond-binary responses were coded as 1 and binary responses as 0.

Procedure

Participants were seated in a quiet room and carried out the experiment on a computer. Each trial consisted of a face accompanied by the question "How would you gender categorize this person?". Each person completed a total of 126 trials. Following Participants were randomly allocated into one of the three response options conditions: binary categories, multiple categories and free text. In the binary categories condition, the only option to respond was "woman" and "man". In the multiple categories condition, this was expanded to include the options "other" and "I don't know". Lastly, the free text condition consisted of an open text box. Participants completed all faces in turn, then filled out answered the gender binary beliefs scale.

Results

RQ1: Do people use non-binary options when they have them?

We used R (Version 4.2.2; R Core Team, 2022) and the R-packages *bayesplot* (Version 1.10.0; Gabry et al., 2019), *brms* (Version 2.18.0; Bürkner, 2017, 2018, 2021), *dplyr* (Version 1.0.10; Wickham et al., 2022), *gcookbook* (Version 2.0; Chang, 2018), *ggplot2* (Version 3.4.0; Wickham, 2016), *papaja* (Version 0.1.1; Aust & Barth, 2022), *Rcpp* (Eddelbuettel & Balamuta, 2018; Version 1.0.9; Eddelbuettel & François, 2011), *tidybayes* (Version 3.0.2; Kay, 2022), *tidyr* (Version 1.2.1; Wickham & Girlich, 2022), and *tinylabels* (Version 0.2.3; Barth, 2022).

To answer RQ1, we first examined the raw distribution of categorizations, presented in Figure 1. From Figure 1 looks like Free text condition largely resembles the binary condition and that furthermore that participants do use use the beyond-binary options in the multiple categories condition.

to do: fix the bug that is producing those ugly red lines at the bottom of this figure

To further test the strength of the evidence, the data from the Free Text and

Multiple Categories conditions were fit to a series of statistical models. For full model specification (including priors) and diagnostics, see the supplementary material. All models were Bayesian mixed effects models with varying intercepts for participants and varying slopes for trials.

The first model was the Null model which included no additional predictors. The second was the Main Effects model which included unique intercepts for each pronoun condition and an overall effect of morph level. Lastly, the Interaction model included unique intercepts as well as unique slopes of morph level for each condition. For a detailed discussion of why this specification is preferred over the traditional dummy-variable approach, see McElreath (2020), but in short it ensures that the priors for each condition are the same, which is necessary for calculating Bayes Factors.

The models were compared using Leave-One-Out cross validation (Vehtari et al., 2017), a method for estimating a model's performance on out-of-sample data. This method of analyses produces LOO values which are not very informative of themselves, but when comparing models, lower values can be determined to show better predictive power. The results of model comparison are presented in Table 1. Table 1 suggests that the Interaction model is the most predictive, but the absolute differnce between the Interaction model and the Main effects model is small and more importantly, the difference is small in relation to the standard error of the difference. This suggests that the data is inconclusive about which model is most suitable. However, to test the specific question raised in the research question 1, we still carried on with the Interaction model.

The estimates of the modelling are visualized in Figure 2. This again suggests that Three specific contrasts were tested with Bayes Factors calculated using the Savage-Dickey Density Ratio (ref). First, whether participants overall made more beyond-binary categorizations in the multiple categories condition than in the free text condition. The evidence suggests fairly convincingly that this is the case (Estimate = 0.02, CI =[0.00],

[0.18], BF_{10}= 97.67). Additionally, based on the curve in Figure 2, we explored whether the evidence supported this diference at morph level 50. The evidence was in favor of this differnce (Estimate = 0.02, CI =[0.00], [0.20], BF_{10}= 17). Lastly, we tested the difference using quadratic weights, though here the difference was inconclusive (Estimate = 0.45, CI = [0.31] - [0.61], BF_{10}= 0.53). *I'm not sure how to interpret this last finding.*

Overall, though, the evidence suggests at least somewhat strongly that when participants have the option of using beyond-binary response options, they use them.

RQ2: Which categories replace the non-binary options?

Based on the shape of Figure 1 it appears that "man" categorizations that are being crowded out by the beyond-binary options. To test whether this was actually the case, we carried out statistical analyses similar to the previous section, again using a mixed-effects model with random intercepts for participants and faces. To explore RQ2, we created three models, a Null model, a Main Effects model and an Interaction model. Similarly, these were then compared using LOO-CV. The results of this are presented in Table 2. This suggests that Interaction model is not the most predictive model, in fact it is the worst. Though, here again, we note that the standard error is quite high, suggesting the proper interpetation is rather that each model is roughly equally as predictive.

Based on the pattern in Figure 1 we did specifically test the contrast between the multiple categories condition and the other two conditions. The evidence were slightly in favor of there being no difference between the multiple categories and the free text conditions (Estimate = -0.49, CI =[-1.23], [0.25], BF_{01}= 4.79) and moderately in favor of no difference between multiople categories and binary categories conditions (Estimate = -0.27, CI =[-0.99], [0.45], BF_{01}= 9.08)

Discussion

To be filled with cogent points.

Experiment 2

Overview

The purpose of experiment 2 was primarily to test categorical perception. If categorical perception occurs, we would expect that scores of femininity to be lower than the percentage of femininity in the faces. Furthermore, if response options change perceptions of gender as a category, we would expect there to be less categorical perception in the multiple categories option.

Method

Participants

Participants ($N = 49$) were speakers recruited through advertising online and on the university campus (M_{age}= 36.67, $SD_{\text{age}} = 12.54$). All participants were informed that participation was voluntary. In term of gender X women and Y men participated The participants were randomly allocated to conditions.

Stimuli & Procedure

The stimuli and procedure for experiment 2 were identical to experiment 1. Experiment 2 differed only the response options conditions. For experiment 2, there response option conditions consisted of single dimension, which ranged from “woman” to “man” and “multiple dimension” which ranged from “not woman” to “woman” and “not man” to “man”. For the multiple dimensions condition, participants rated the same faces according to both scales, but on separate trials.

Results

The mean ratings in both conditions are presented in Figure 3.

As a further test of the research question, we also fitted the data to a Bayesian mixed effects model, with participants and faces modeled as random intercepts.

Additionally, the morph levels were entered as categorical predictors, rather than as continuous variables. Similar to study 1, the initial approach consisted of several models which were compared against each other using LOO-CV. Again, the models were a Null model, with no additional predictors, a Main Effects model with main effects of morph level and condition, but no interaction, and a Interaction model (for complete model specification, see the Supplementary material.)

Based on the results of LOO-CV, we continued with the Interaction model. We carried out two comparisions. The first was a quadratic contrasts *which I have still to carry out*. Because the critical levels where we might expect to see a differ, we also compared the mean rating at 33.33 morph and at 66.67 morph. At 33.33 the evidence strongly suggested that the two conditions are the same (Estimate = 0.28, CI =[-3.91], [4.51], BF_{01}= 31.57). This was also the case at 66.67 (Estimate = 2.29, CI =[-2.03], [6.57], BF_{01}= 19.17). Overall, both conditions showed fairly strong tendencies toward categorical perception and they did not differ in this regard.

Discussion

Overall discussion

References

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://github.com/crsh/papaja

Barth, M. (2022). *tinylabels: Lightweight variable labels*. https://cran.r-project.org/package=tinylabels

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Chang, W. (2018). *Gcookbook: Data for "r graphics cookbook"*. https://CRAN.R-project.org/package=gcookbook

Eddelbuettel, D., & Balamuta, J. J. (2018). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *The American Statistician*, *72*(1), 28–36. https://doi.org/10.1080/00031305.2017.1375990

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in bayesian workflow. *J. R. Stat. Soc. A*, *182*, 389–402. https://doi.org/10.1111/rssa.12378

Kay, M. (2022). *tidybayes: Tidy data and geoms for Bayesian models*. https://doi.org/10.5281/zenodo.1308151

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Wickham, H., & Girlich, M. (2022). *Tidyr: Tidy messy data*. https://CRAN.R-project.org/package=tidyr

Table 1

Relative predictive power of models describing the outcome on the categorization task

	LOO diff	St. Error diff	LOO	St. Error LOO
Interaction	0.00	0.00	-234.17	23.23
Main Effect	-2.46	2.71	-236.63	23.07
Null	-18.83	6.02	-253.00	24.51

Note. LOO diff refers to the difference in loo between the model and the most predictive model. The first row describes the most predictive model, which is why the difference is 0

Table 2

Relative predictive power of models describing the outcome on the categorization task

	LOO difference	St. Error diff	LOO	St. Error LOO
morph_only	0.00	0.00	-1343.71	43.17
main_effects	-1.85	0.86	-1345.56	43.17
condition_only	-4.98	5.26	-1348.69	44.49
Null	-5.36	4.88	-1349.07	44.12
interaction	-6.69	3.02	-1350.40	43.48

Note. LOO diff refers to the difference in loo between the model and the most predictive model. The first row describes the most predictive model, which is why the difference is 0

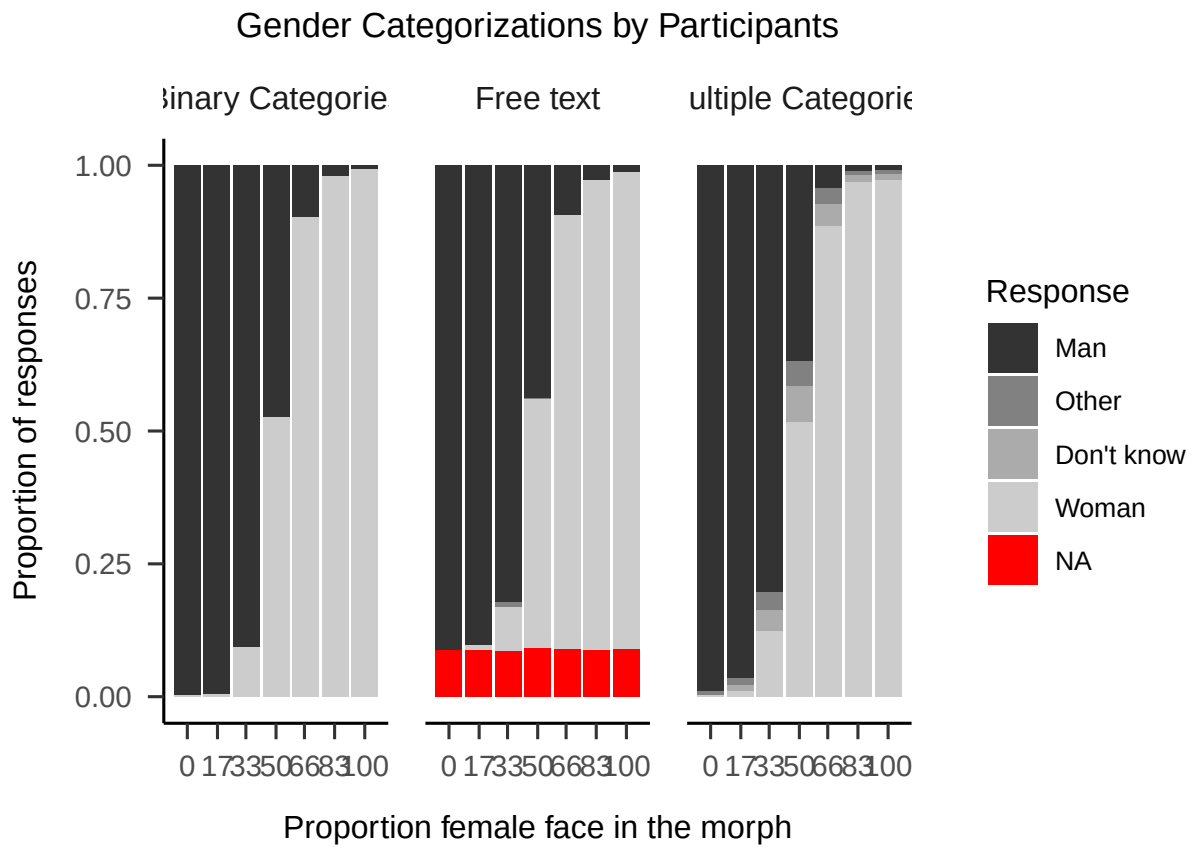

Figure 1

Gender Categorizations by Participants

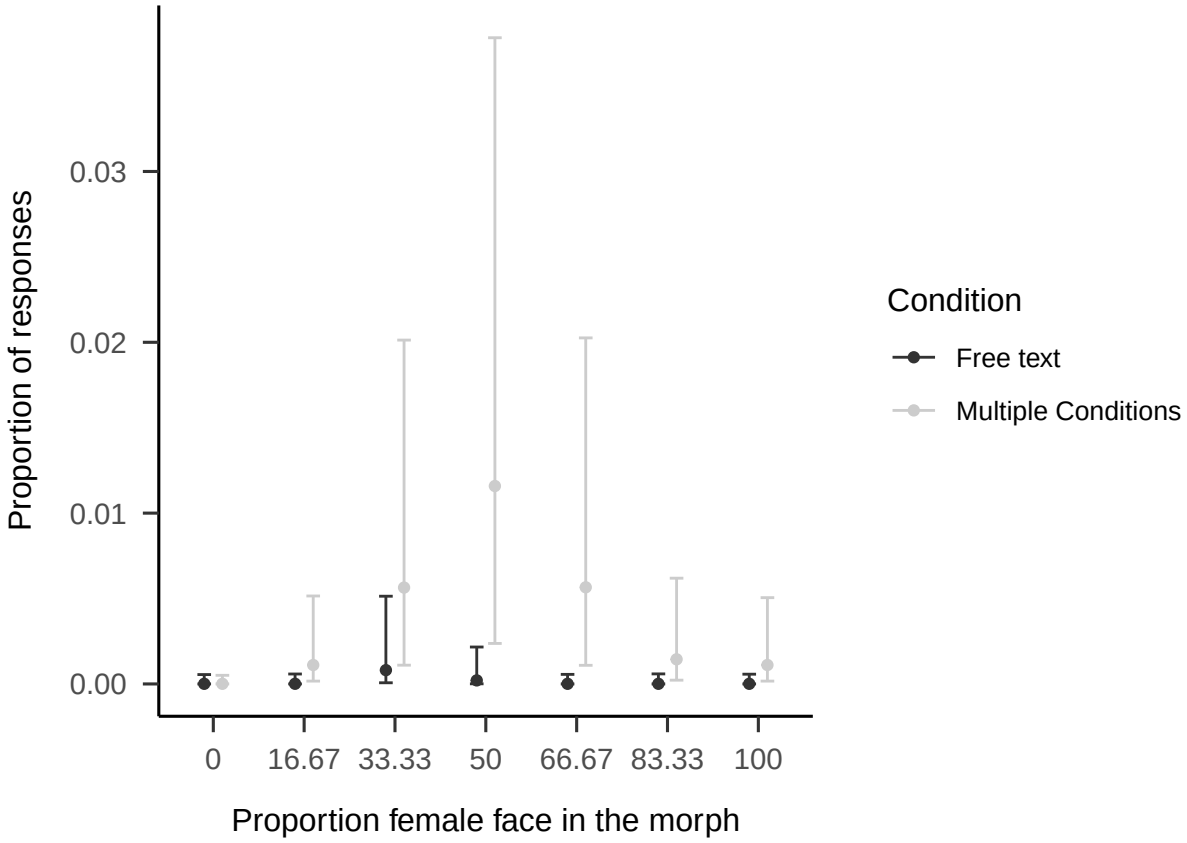

Figure 2

Proportion of beyond-binary responses in the Mulitple categoreies and Free Text conditions

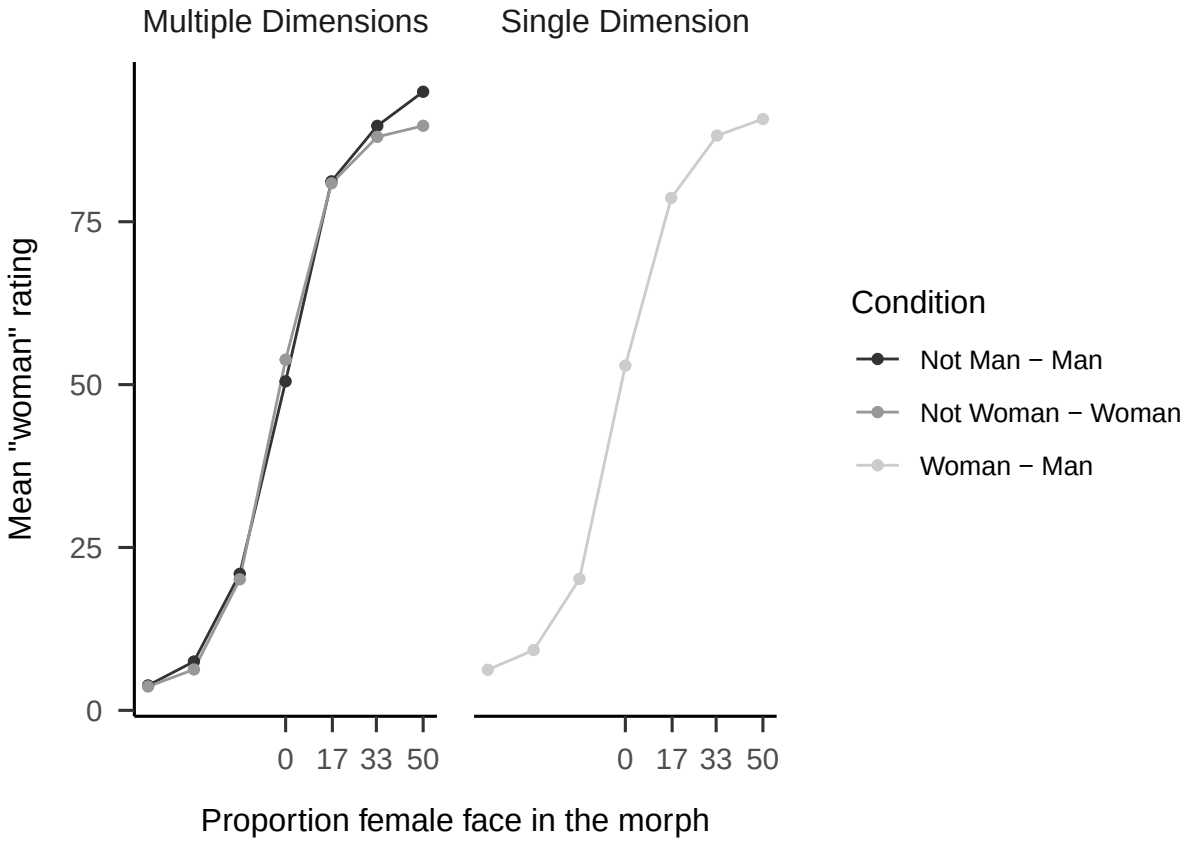

Figure 3

Mean ratings of faces in Single dimension and multiple dimensions

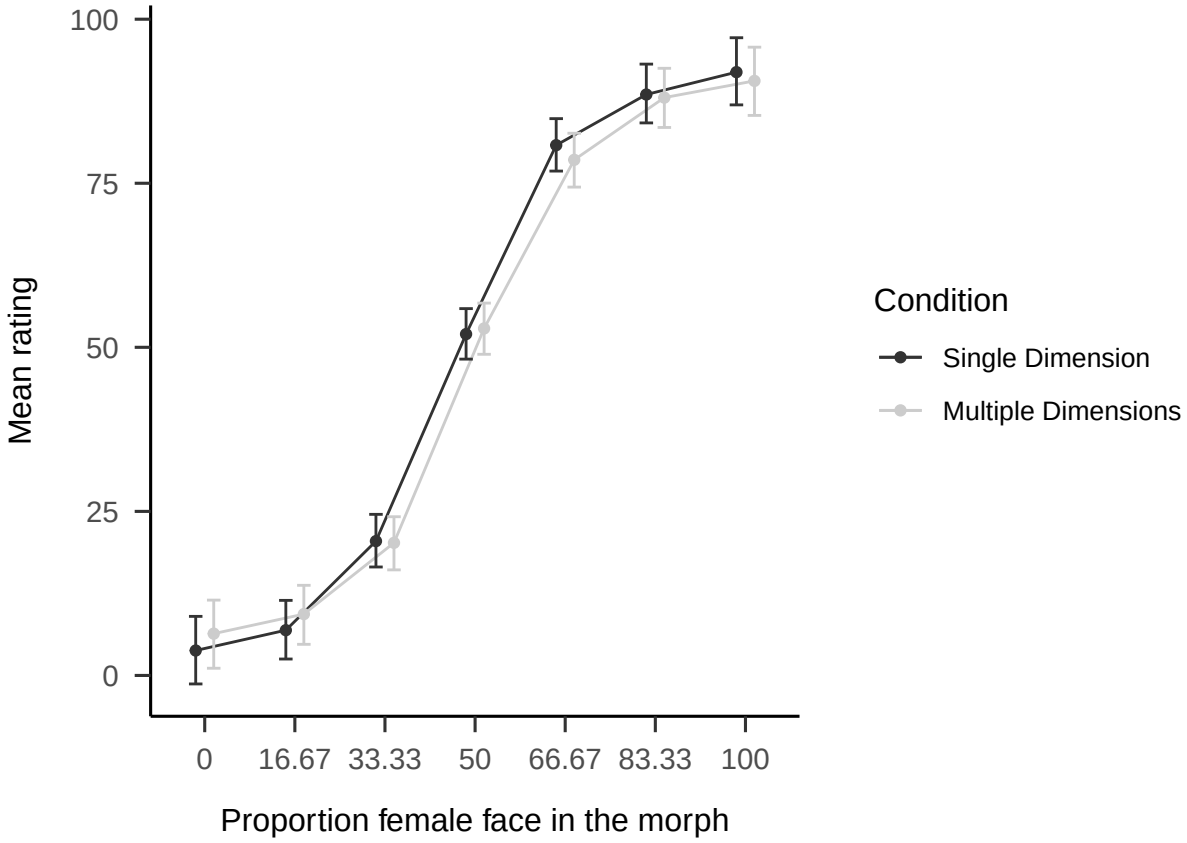

Figure 4

Mean gender ratings in Single Dimension and Multiple Dimensions conditions