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 declare no conflict of interest.

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)

Precision is key when measuring constructs in psychological research. One domain where precision is lacking is in the use of binary response options for gender, which fails to capture the complex and fluid nature of gender/sex. The gender binary is most clear in terms of legal gender, where the alternatives in most countries are confined to women or men, this dichotomy does not accurately reflect the diverse biological and self-perceived gender identities that exist (Lindqvist et al., 2019; Hyde et al., 2019). While more flexible options for self-definition of gender identity has become more common in research, the use of non-binary gender options in stimuli and participant response measures remains limited. In light of this, the current study aims to highlight non-binary alternatives to measuring gender categorization and investigate how gender perception is influenced by such non-binary options.

Alternative approaches to binary self-classification of gender have a deep history in psychology. For example, Bem (1974) developed scales to measure femininity and masculinity as separate traits, finding that many individuals exhibit a mixture of feminine and masculine traits. More recently, Joel and colleagues (2014) not only asked participants which gender category they identified with, but also whether they had ever experienced a different gender identity. They found that approximately 30% of respondents had experienced an alternative gender identity at some point in their lives. Additionally, several guides have been developed that recommend the use of open textboxes or multiple options when asking participants to indicate their gender. When participants are given such textboxes, they frequently used them. Thes studies suggest not only that alternatives are possible, but also that participants will use them when they have the chance, and potentially, the possibilities around gender change.

In contrast, the literature on gender categorization often treats gender as a binary category. Gender categorization is a cognitive process that occurs when individuals perceive others (ref). Researchers in this field have explored the speed and automaticity of gender perception in faces, as well as which facial features are associated with specific gender categories, such as women and men (ref). Generally, the findings indicate that gender is rapidly and automatically categorized, with facial features such as skin smoothness, jawline, and hair length used to determine gender identity. Lastly, studies in this field have indicated that people percieve faces categorically (Campanella et al., 2001). In other words, However, these studies typically do not address the complex nature of gender or consider alternative response options (ref).

Instead, gender categorization is most often measured through a forced-choice task in which participants are forced to indicate either "female" or "male" when presented with a face (see, for example, Cloutier et al., 2005; Campanella et al., 2001; Webster et al., 2004; Zhao & Bentin, 2008). A slightly different approach asks participants to rate faces on a gender scale as a quality, often using "feminine" and "masculine" as endpoints on a single scale (e.g., D'Ascenzo et al., 2015; others). Despite some variations, therefore literature overwhelmingly presents gender as a binary in studies of gender categorization.

The use of binary gender measures presents significant problems in accurately reflecting the complexity of gender. Such measures reinforce the notion of gender as a binary concept, thereby invalidating non-binary genders. This not only misrepresents the reality of gender but also perpetuates discriminatory attitudes towards non-binary individuals. Consequently, it is imperative to consider alternative measures that acknowledge and respect the diversity of gender identities.

Overview of planned studies

Based on the problems with binary measurement, we planned two studies to demonstrate alternatives to binary measurment of gender categorization and to investigate how such alternative impact participants results. The purpose of bothg experiment was to answer the following three research quetions.

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

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

Research question 3: Can response options which do not present the categories of woman and man as oppositional reduce categorical perception.

Experiment 1

The purpose of study 1 was to test research questions 1 and 2. Therefore, the

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 the gender binary beliefs scale.

Data analysis

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). The data were handled both descriptively and fit to Bayesian mixed-effects models. To answer each research question, we used a two-step approach which began with a model comparison approach followed by Bayes factor tests of specific contrasts. In all cases, the models included varying intercepts for both participants trials.

RQ1. In research question one, we investigated whether XXX

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.

RQ2. In research question two, we investigated whether XXX

Results

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

The raw distribution of categorizations is presented in Figure 1.

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

As described in the methods, to investigating RQ1 we fit a Null Model, a Main Effects Model and and interaction model to the data. 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, but both are superior to the Null model. As model comparison did not conclusively discount the Interaction model, we continued by testing specific, relevant contrasts.

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

The results from experiment 1 suggest that some participants do use the beyond-binary options when they have them, however only when these are explicitly spelled out. When participants are implicitly able to enter whatever they like, most still fell back on using woman/man. Furthermore, the results suggests that overall, even when participants used the beyond-binary options, this did not systematically affect their overall pattern of responses in terms of woman and man categorizations.

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_{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

Experiment 2 was designed to test whether response options which did not present women and men as opposing categories changed participants categorical perception. Overall, the results indicated that in terms of categorical perception, the two conditions were very similar, suggesting that the binary view of language is very strong and that participants do not change their view of gender depending on

Overall discussion

Overall, some important findings emerged from this experiment. First, there was strong evidence that participants will use beyond-binary actions to categorize faces, if given the options. It is also clear that providing these options do not change the overall distributions of scores

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

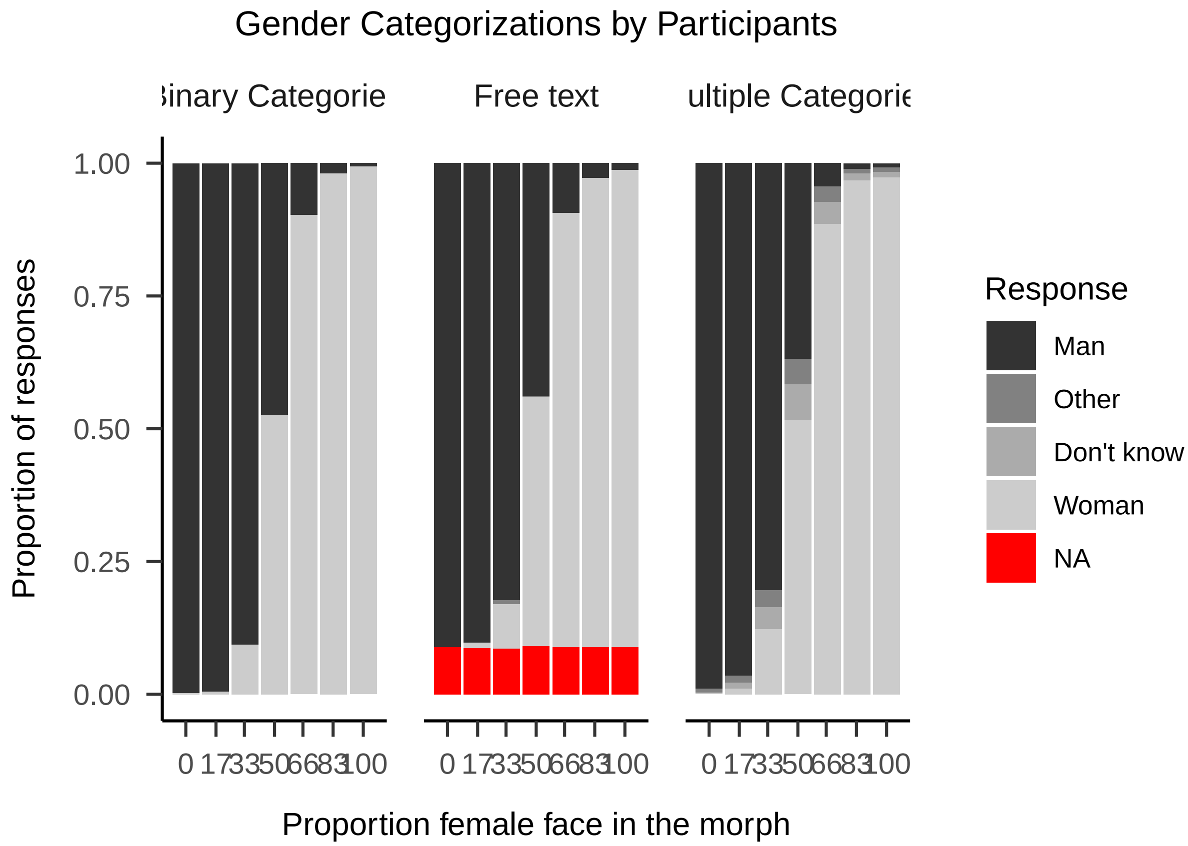

Figure 1

Gender Categorizations by Participants

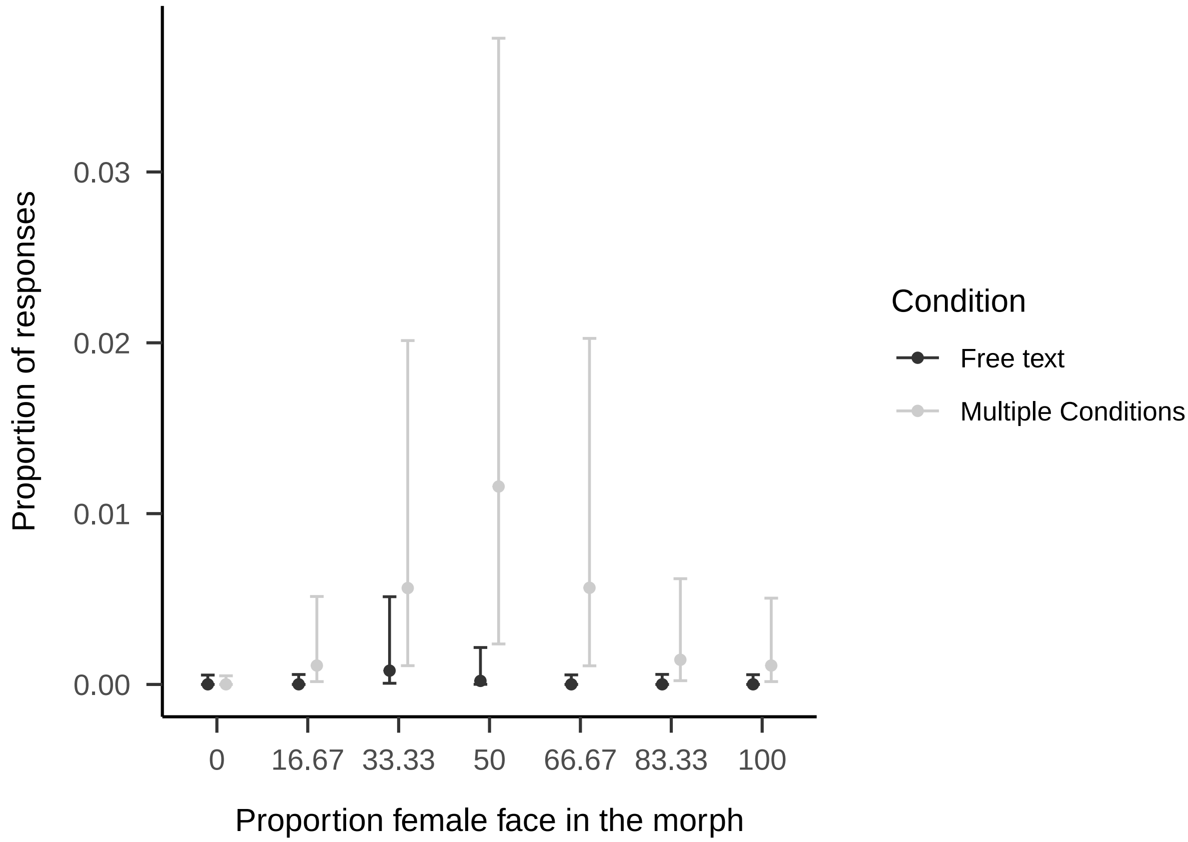

Figure 2

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

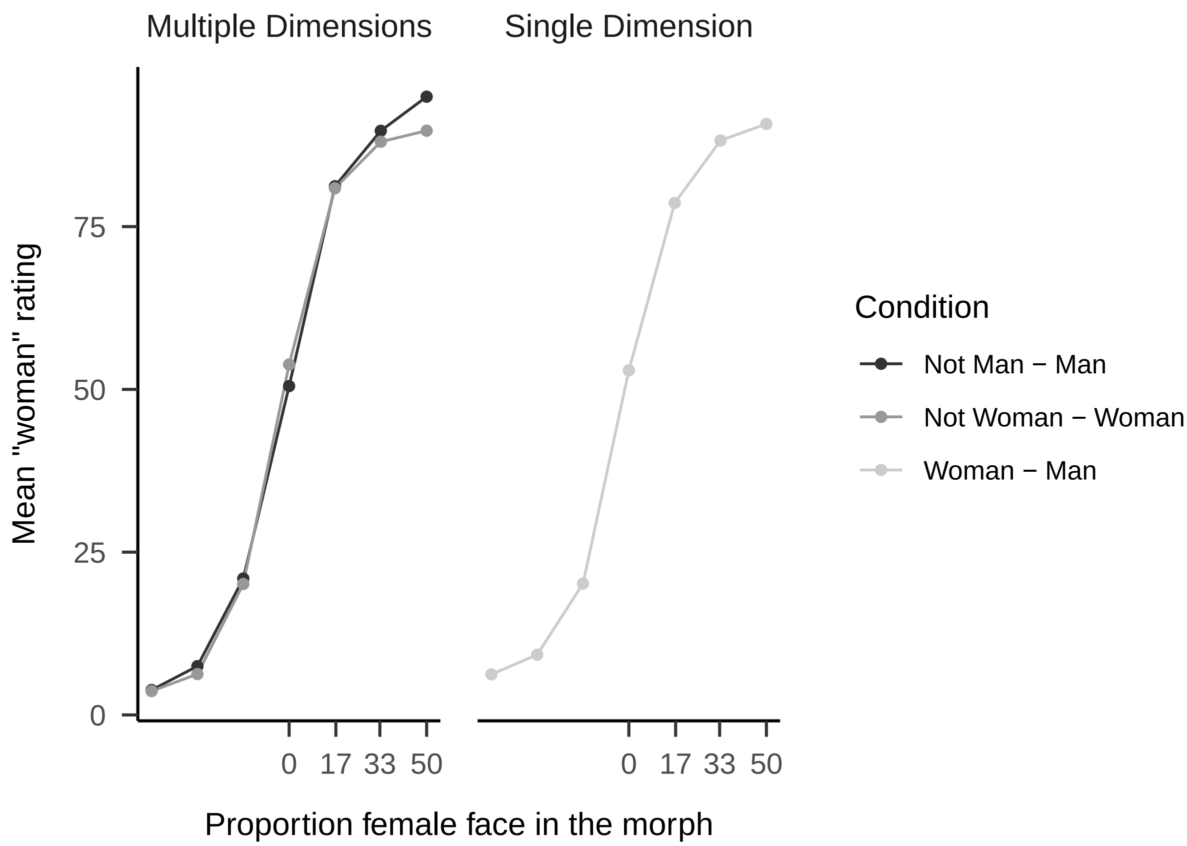

Figure 3

Mean ratings of faces in Single dimension and multiple dimensions

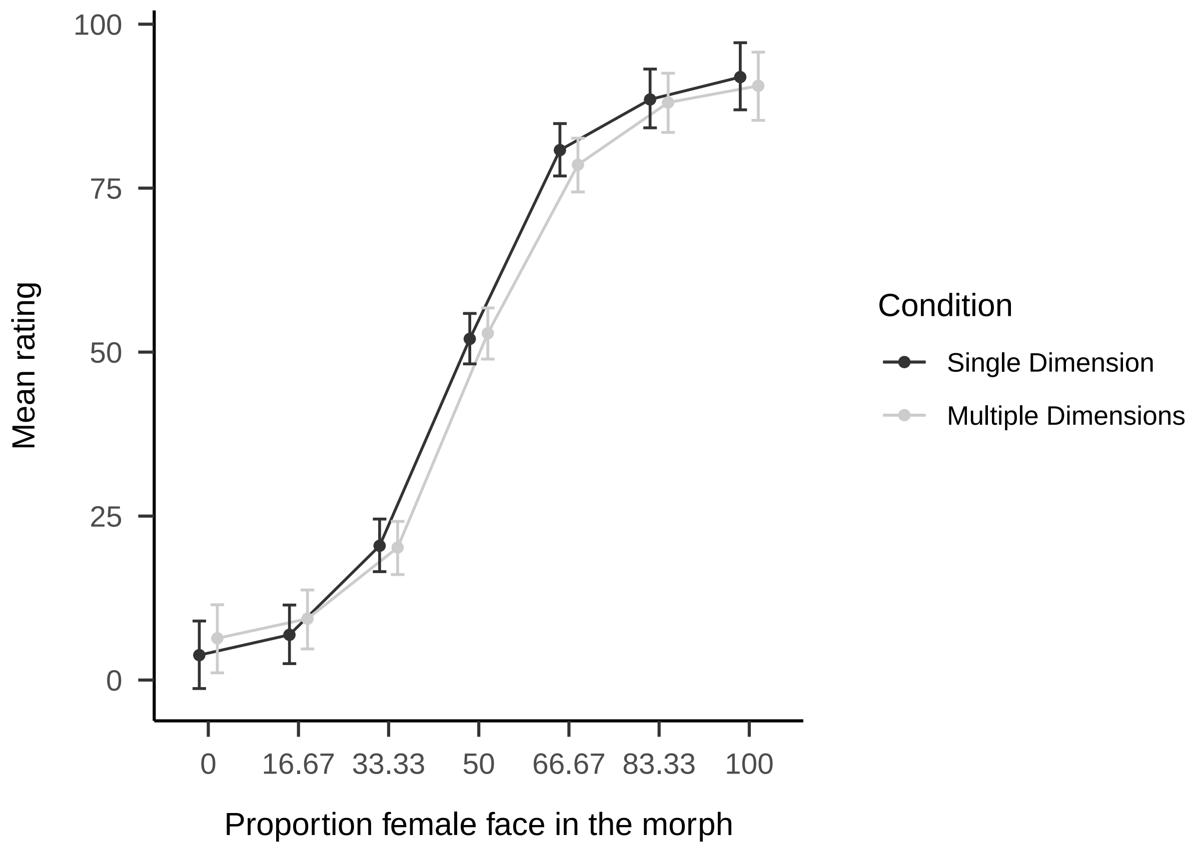

Figure 4

Mean gender ratings in Single Dimension and Multiple Dimensions conditions