# Toward Inclusive Research: The Effect of Response Options on Gender Categorization of Faces

Elli van Berlekom, Stefan Wiens, and Marie Gustafsson Sendén

Psychology Department, Stockholm University

# Author Note

Elli van Berlekom  http://orcid.org/0000-0002-1949-5600

Stefan Wiens  http://orcid.org/0000-0003-4531-4313

Marie Gustafsson Sendén  http://orcid.org/0000-0002-8393-5316

Correspondence concerning this article should be addressed to Elli van Berlekom, Psychology Department, Stockholm University, Email: elli.vanberlekom@psychology.su.se

# Abstract

Psychological research often treats gender as a binary and unidimensional even though this oversimplifies the variability of gender. In two experiments, we studied how alternative response options influence how people perceive the gender of a racially diverse set of morphed faces. Experiment 1 (N = 71), compared one-dimensional and two-dimensional scales for gender categorization. Results indicated that participants consistently used the ends of dimensional scales, resulting in highly categorical perceptions. Experiment 2 (N = 105), compared traditional binary response options with multiple categories and free-text answers. Results indicated that providing more response options than the binary (e.g., “non-binary”, “I don’t know”) made participant perceive gender as more diverse, whereas free-text options did not change a binary categorical perception of gender. Despite the opportunity to categorize faces beyond the binary, the predominant categorizations remained as ‘woman’ or ‘man’. We conclude that multiple response options seems to be the best way to increase perceivers understanding of gender beyond the binary.

# Toward Inclusive Research: The Effect of Response Options on Gender Categorization of Faces

Many transgender and gender diverse people experience gender as fluid, diffuse, and outside the typical binary of women and men (Hyde et al., 2019; Richards et al., 2016a). Unlike cisgender people - who identify with their assigned gender at birth - transgender people identify with a gender different from their assigned sex at birth (Levitt & Ippolito, 2014). Moreover, many transgender people identify as nonbinary, which can be either an identity in and of itself or an umbrella term for a wide variety of gender identities other than woman or man (e.g., genderqueer, agender, genderfluid) (Monro, 2019).

Surveys and questionnaires that collect data on gender have typically used response formats that do not account for this gender diversity, for example by including only the categories woman/female and male/man as response options (Saperstein & Westbrook, 2021). This is beginning to change; it is increasingly common for studies to include response formats that are sensitive to a wide variety of gender identities, at least when collecting data on participants’ own gender identities (Hyde et al., 2019). Research on gender categorization of others, however, is still dominated by binary response options (e.g., Campanella et al., 2001; Habibi & Khurana, 2012; Jung et al., 2019). In this study, we aim to challenge this trend and test how response formats influence participants’ gender categorizations.

## Gender and Gender Categorization

[what is gender] To understand why response formats matter, it is important to understand the concept of gender itself. Gender is a complicated and multifaceted concept with both personal and social aspects (Hyde et al., 2019). The personal aspects of gender include a person’s physical body, how they choose to present themselves as well as their internal sense of gender (i.e. their gender identity). The social aspects of gender include beliefs about which gender categories exist, prescriptions and proscriptions for appropriate behavior for different gender categories, as well as the language available to refer to gender, including gendered pronouns and the names of gender categories (Lonergan & Palomares, 2020). The personal and social facets of gender are interconnected (Morgenroth & Ryan, 2021); social norms constrain the personal aspects of gender and, inversely, personal enactments of gender may influence the social norms over time (Butler, 1999). For example, Western societies have long constructed gender as a binary, with language and media rendering transgender and non-binary gender invisible, limiting gender diverse people’s ability to know and express their identities (Richards et al., 2016b). Inversely, as gender diverse people become increasingly visible, norms and language terms are changing, and the wider conceptions of gender change to incorporate transgender and non-binary gender categories (Monro, 2019, fasoli), though not without resistance (Vergoossen, Morgenroth, Tee & Hegarty)

[gender categorization - intro] Gender categorization is the act of ascribing gender categories to others and is influenced by both the personal and social aspects of gender (Kessler & McKenna, 1978). For example, people base their gender categorizations on other’s external features (Yamaguchi et al, 2013, Cloutier et al., 2014), and which features are associated with which gender identities is partially a result of norms (Butler, 1999).

The processes that shape gender categorization are an important object of study because gender categorization may activate stereotypes and influence judgements of other traits and characteristics, such as aggression (Freeman & Anbady, 2011). Moreover, many studies suggest that gender categorization is a central and important feature of social interactions (Kessler & McKenna, 1978, Liberman et al., 2017); Gender categorization is easy (Yang & Dunham, 2019), fast (Tomilleri & Castelli, 2012), and does not require much visual input (ref). Some have gone so far asto conclude that gender categorization is an automatic (ref) and unavoidable feature of social interactions (Fiske, 1998).

[categorical perception] Additionally, it has been argued that gender is perceived categorically (Campanella et al., 2001). This is supported by research showing that when people categorize faces that vary on a continuum feminine to masculine, their categorizations are exaggerated compared to the level of facial gender (Rule et al., 2012; Campanella et al., 2001; Atwood et al., 2024). For example, a 60% female face was categorized as a woman by closer to 80% of participants (Campanella et al., 2001). Such categorical perception effects have been observed for other domains, such as colors (blue, green) and phonemes (“ga”, “ba”) and suggest that people view the stimuli as consisting of two clear groups or categories (Simanova et al., 2016). The observation of categorical perception of gender, therefore, implies that people treat gender as a binary consisting of women and men only.

## Gender categorization beyond the binary

[Binary gender categorization is not inevitable] Several recent studies suggest that this binary view of gender may be learnt. For example, in one study, participants were tasked with keeping track of three people carrying out a conversation [Gallagher et al., 2024]. On both implicit and explicit indices of gender categorization, transgender status, age and views on gender predicted participants’ likelihood of engaging in automatic gender categorization, suggesting that prior experience played a major role [Gallagher et al., 2024]. Additionally, two studies have shown that categorical perception of gender can be disrupted. Hegarty et al. (year) disrupted categorical perception by presenting faces only in the left visual field. These stimuli were processed by the right brain hemisphere, implying that only the left brain hemisphere displayed categorical perception of gender. The categorical perception effect was also disrupted by the introduction of a third, intermediate response option [Atwood et al., 2024]. These studies together, suggest that the binary view of gender is learned, which implies that it can be unlearned. (this paragraph needs a little more juice)

[Gender categorization research does not account for gender diversity] Moreover, research gender categorization has largely assumed that gender is binaryLindqvist et al. (2020) often expressed through operationalizations of gender categorization through binary response formats. Two examples of such formats are gender categorization expressed as a single dimension (feminine-masculine) and as binary response options(woman/female and man/male). Both response formats imply ideas about gender: a single dimension suggests that femininity and masculinity are mutually exclusive opposites; and binary response options suggest that those are the only gender categories that exist. Given the findings that gender categorization is susceptible to influence from short-term, methodological manipulations, it is possible that binary response formats bias participants toward a binary conception of gender.

[Gender categorization beyond the binary is becoming a topic of study on its own merits.] Another consequence of this focus on binary gender categorization is that not much is known about categorization of non-binary gender. The increasing prominence and ongoing marginalisation of nonbinary people (Monro, 2019; Jacobsen, 2023) means that this is an important topic of research. For example, there is a misalignment between how many categorize gender, and the existing gender diversity; many people’s tendency to treat gender as both binary and synonymous with sex leads them to misattribute gender (Howansky, 2019). Additionally, it is unclear whether transgender and non-binary people are categorized and stereotyped in the same way, or whether these groups face unique stereotypes. Studies are beginning to answer these questions and more related to categorization of transgender and non-binary gender (see Atwood et al., 2024; van Berlekom et al., 2024; Weissflog et al., 2023 for some recent examples) but this area is still very much nascent. As more research begins to investigate gender categorization beyond the binary, it becomes increasingly important to develop sensitive response formats that accurately represent non-binary gender without implicitly reinforcing gender hierarchies.

## Response Formats Beyond the Binary

*I have collapsed the research of self and other categorization. Why did I do this? Basically because I found that the paper was getting quite repetitious* Over the years, a number of studies have investigated gender categorization using response formats that challenge gender hierarchies. One solution is to conceptualise femininity and masculinity as independent entities. Unlike one-dimensional treatments of gender, such a two-dimensional approach does not imply that femininity and masculinity are opposites and that an increase in one necessarily entails a decrease in the other. One of the earliest examples of scales to use such an approach was the Bem Sex Role Inventory (BSRI), a questionnaire scale which aimed to measure gender as a psychological trait (Bem, 1974). When this scale was used, some participants exhibited clear femininity or masculinity but many scored as “androgynous” (high femininity, high masculinity) and “neutral” (low femininity, low masculinity), destabilising binary notions of gender (Liben & Bigler, 2017). Such a treatment of gender has been recommended both for participants categorizing their own gender (saperstein) used in face categorization research (Wittlin, 2018).

Another approach has been to broaden the response options when participants are selecting from an established set of categories. Saperstein and Westbrook (2021) suggested that surveys measuring gender include a range of response options, such as non-binary, other, trans man, agender, and more. This approach is becoming quite common for research that measures participants’ categorization of their own gender (li et al., Saperstein et al.). Indeed, variations of this task have also been used to study non-binary categorizations (van Berleko et al., Weissflog et al). One issue with this task is that it does not challenge the idea that gender can be visible from appearance, and may therefore implicitly reinforce the stereotype that non-binary people are androgynous.

Finally, Lindqvist et al. (2020) suggested that the most sensitive measure for self-categorization would be an open text entry where participants can fill in their gender in an open-ended format. The free text response has the advantage of being completely unconstrained, allowing participants to enter any category, including categories which may not have occurred to the researchers. Moreover, the acceptable terms sometimes shift over time as more marginalised voices are heard. This method has, as far as we are aware, not been tested in the context of categorization of others.

## The Present Research

Study 1 thested the influence of one and two-dimensional response options on categorical perception. We compare gender categorization measured using one-dimensional response options (ranging from woman to man) and two-dimensional response options (ranging from woman/man to not woman/not man). If categorical perception occurs this would suggest that participants treat gender as consisting of only two categories: women and men. Accordingly, a reduction in this effect would suggest participants take a more fluid view of gender. We tested two research questions:“Do participants respond categorically to faces?” (Research Question 1) and “Does a one-dimensional rating format elicit weaker categorical responses than a two-dimensional rating format (Research Question 2)?”

Study 2 investigated gender categorization using non-binary gender using multiple response options or free text answers. In Study 2 we were mainly interested in how these two measures affect gender categorization in comparisons to only binary response options (Research Question 3). On inspecting the data, we noted a pattern that seemed to indicate that non-binary responses crowded out male responses. Therefore, study 2 also investigated the categorization of women and men (Research Question 4).

# Study 1

## Method

### Participants

Swedish participants (*N* = 71) took part in the Stockholm University campus (*M*age= 37.87, *SD*age = 14.08, Range = 18 - 73). Participants included 33 women, 35 men, and 2 participants who did not indicate gender (self-identified gender was measured using an open-ended text box, following Lindqvist et al., 2020). Participants were randomly allocated ~~into~~ to one of the two response option conditions (*N*control = 33, *N*experimental = 38). Participants were monetarily compensated for their time (100 sek). In accordance with the Helsinki Declaration, all participants were informed that participation was voluntary and gave written consent to participate in the study.

### Stimuli

The experiment included Black, Asian, and White faces from the London Face Database (DeBruine & Jones, 2017) and the Chicago Face Database(Ma et al., 2015) morphed with Webmorph (DeBruine, 2018). We selected matched pairs of faces of women and men, ensuring that the women were rated at similar levels of feminine as the men were rated masculine ~~using the codebook provided by the researchers~~ according to the norming data (Ma et al., 2015). The morphs were made in 7 steps, from completely feminine to completely masculine. We defined facial gender as the degree of the female face present in the morph. In other words, a 33% face was slightly masculine, a 50% face was an even mixture of the two faces, and a 100% face consisted only of the woman’s face. Because there were 18 face pairs morphed in 7 steps, the total number of faces was 126.

### Procedure

Participants completed the experiment on a computer in a quiet room. Each trial consisted of a face accompanied by the question, “How would you gender categorize this person?”. In the *one-dimensional* condition, participants rated gender based on a single continuum, with the anchors marked *woman* and *man*. In the *two-dimensional* condition, participants rated each face once on a *woman* continuum (the anchors were marked *not woman* and *woman*) and once on a *man* continuum (anchors were marked *not man* and *man*). The separate continua were presented on different trials, and the order of trials was completely randomized in both conditions (see [Figure 2](#fig-exp2-trial)).

### Data analysis

We used R (Version 4.2.2; R Core Team, 2022) and the R-packages *brms* (Version 2.18.0; Bürkner, 2017, 2018, 2021), *papaja* (Version 0.1.1; Aust & Barth, 2022), and *tidyverse* (Version 1.3.2; Wickham et al., 2019). We fit Bayesian mixed-effects models to the data to test for patterns of responses consistent with categorical perception. In all models, facial gender (0 to 100 in seven steps) and response options (one-dimensional, two-dimensional) were included as fixed effects. Additionally, all models included varying intercepts for both participants and faces and varying slopes for facial gender. We modeled the predictor facial gender as an ordered factor with seven levels corresponding to each of the seven morphing steps. This allowed us to test for the kinds of non-linear patterns that would be observed under categorical perception, where changes in facial gender would be expected to have a larger effect on rated gender near the midpoint than at extreme values.

## Results

We examined the relationship between ratings of woman and man in the two-dimensional condition. These were highly negatively correlated (R = -0.86). Therefore, man ratings in the multiple dimensions were reverse-coded for subsequent analyses. Second, we examined whether participants responded categorically to faces (Research Question 1). Individual-level (thin lines) and group mean (thick lines) responses are visualized in [Figure 3](#fig-desc-two). If participants respond according to the morph level, the lines should be a straight diagonal. Instead, [Figure 3](#fig-desc-two) shows that most participants display a non-linear S-shape, and this was also the pattern of the group means. Note that in the two-dimensional condition, participants rated each face twice.

To further test whether the faces were rated categorically, we calculated the difference between the mean ratings when facial gender was 33% and 67%. If participants respond linearly, this difference should be 34. Instead, in both the one-dimensional condition (*M*1D = 59.58, CI1D = [53.65, 65.26]) and the two-dimensional condition (*M*2D = 58.75, CI2D = [52.53, 65.08]) this difference far exceeded 34 and the narrow credible intervals suggest these measures were precisely estimated. We interpret this to mean that participants responded categorically.

Finally, we tested whether the categorical perception was reduced in the two-dimension condition compared to the one-dimension condition (Research Question 2). In other words, we calculated the mean difference between 67% faces and 33% faces ~~across the two~~ in both conditions. The results suggested that categorical perception was not reduced by two-dimensional response options (Difference = -0.83, CI = [-5.57, 7.24], BF01= 30.47).

## Discussion

Participants responded categorically when rating faces in terms of gender. Additionally, two-dimensional response options did not reduce this effect. Indeed a highly binary view of gender was present and participants treated womanhood and manhood as opposites even though the scale would allow them to be more flexible. However, this scale only implicitly challenged the binary, as no diverse gender options were present.

# Study 2

Study 2 tested a wider range of response options that explicitly challenge the gender binary. These were adapted from common ways to measure participants’ self-categorization of gender (Lindqvist et al., 2020; Saperstein & Westbrook, 2021). In Study 2 we compared three types of response options in a gender categorization task: 1) only woman and man; 2) woman, man and other and 3) an open text box for participants to type in their response. As a control condition, we also included a condition with only woman and man as response options.

## Method

### Participants

Swedish participants (*N* = 100) took part in the study at the Stockholm University campus (*M*age= 36.89, *SD*age = 13.69, Range = 18 - 69). Self-identified gender was measured using an open-ended text box as recommended by (Lindqvist et al., 2020). The final sample included 56 women, 47 men, and 2 participants who did not indicate gender. All participants were informed that participation was voluntary and all gave written consent to participate in the study. Participants were randomly allocated to one of the two response option conditions (*N*binary = 32, *N*multiple = 36, *N*free\_text = 32). Participants were monetarily compensated for their time (100 sek).

### Stimuli

The stimuli were identical to those of Study 1.

### Design and Procedure

The experiment used a between-participants design. There were three conditions with different response options: binary categories, free text, and multiple categories (see [Figure 4](#fig-exp1-trial)). In the *binary categories* condition, the response options consisted of two categories: woman and man. In the *free text* condition, the response options consisted of an open text box. In the *multiple categories* condition, the response options consisted of four categories: woman, man, other, and I don’t know.

Participants completed the experiment on a computer in a quiet room. Each trial consisted of a face accompanied by the question, “How would you gender categorize this person?” After being allocated to one of the three conditions, participants categorized 126 faces according to the response options in their condition.

The outcome was responses to the categorization task. For analysis purposes, ~~these were recorded in the following ways~~ [two new variables were created]{style=“color:green”;}:

*Other categorizations* represented the trials where participants categorized faces as any other category than woman or man. This was computed by dichotomizing the variable so other was coded as 1 and all other responses were coded as 0. In the free text condition, participants’ responses were manually coded so that of other and non-binary counted as other.

*I don’t know* responses represented trials where participants did not categorize any gender category. I don’t know was coded as 1 and all other responses were coded as 0. In the free text condition, participants’ responses were manually coded so that variations of unsure and I don’t know counted as I don’t know.

### Data analysis

Bayesian linear mixed effect models were fit to the data in study 2. These models were the same as in Study 1 apart from the outcome, which was binomial and accordingly had to be modeled as a binomial distribution.}

## Results

[Figure 5](#fig-desc-1). A simple visual inspection of Figure [Figure 5](#fig-desc-1) suggests that most faces were categorized as women or men. Participants did categorize faces outside of this binary in the multiple categories condition, as [Figure 5](#fig-desc-1) shows, and most such categorizations were made in response to androgynous faces.

[Figure 5](#fig-desc-1), however, only illustrates the total number of categorizations across all participants. This obscures the fact that some participants made many categorizations beyond the binary and some made few or none at all. [Figure 6](#fig-desc-nbo) illustrates how many categorizations (y-axis) beyond the binary participants made. Each bar represents how many participants (y-axis) made a certain number of categorizations (x-axis). The different colors denote the different categorizations. Participants who only categorized faces as women or men are not represented in figure [Figure 6](#fig-desc-nbo). In the Free Text condition, only two participants made any other categorization than woman and man, whereas more than half did so in the Multiple Categories condition (see [Figure 6](#fig-desc-nbo) ).The Bayesian mixed effects model suggested that participants made categorizations beyond the binary in the multiple categories condition compared to the free text condition (OR = 5.56, CI =[1.1, 27.97], BF10= 4.55).

An inspection of [Figure 5](#fig-desc-1) suggests that participants made fewer man categorizations in the multiple categories condition than the other two conditions. We tested this by examining only responses of woman or man. [Figure 8](#fig-wtf) illustrates proportions of responses of women and men. Each dot represents a single participant, and the position of the dots on the y axis shows the proportion of faces that participant categorized as man; the boxplots show median and interquartile range proportion of faces categorized as men (in this data set, with categorizations beyond the binary removed, any face not categorized as a woman was categorized as a man). Overall rates of binary categorizations were similar across the three conditions (see [Figure 8](#fig-wtf)).

We treated the binary categories condition as the control against which the other two conditions were compared. The results suggested that the proportion of faces categorized as women was similar in the Multiple Categories and Binary Categories conditions (OR = 0.68, CI =[0.4, 1.17], BF01= 5.98). The results suggested that the proportion of faces categorized as women was the same in the Free text and Binary Categories condition (OR = 1.03, CI =[0.6, 1.78], BF01= 15.27). In sum, neither the free text nor the multiple categories condition changed the pattern of categorization of women and men compared to the binary categories condition.

## Discussion

In Experiment 2, we tested how free text options and multiple categories affected participants’ responses beyond the binary. Some participants made some categorizations beyond the binary in the multiple categories condition, but virtually none did so in the free text condition. Furthermore, additional response options reduced the absolute number of faces categorized as women and men (as participants selected some of the other options) but did systematically reduce categorizations of men more than women or vice versa.

# General Discussion

Across two experiments, we tested how different response options influenced gender categorization. In Study 1, we compared two-dimensional scales with one-dimensional controls. We found that participants responded categorically, and this was the case in both the control condition and the two-dimensional condition. In Study 2, we compared free text and multiple categories. We found that only multiple categories elicited beyond-binary responses. Compared to binary control, neither changed the pattern of categorizations of women and men.

The results from Study 1 are consistent with previous work on categorical perception of gender in faces (Campanella et al., 2001, 2003). Participants exhibited a categorical pattern of responses where ratings of gender were more extreme than the facial gender. This implies that participants had a conception of gender as consisting of two distinct categories. Furthermore, the two-dimensional ratings did not reduce the strength of the categorical effect. This suggests that, at least in the present sample, two-dimensional response options were not enough to reduce the binary gender norms.

This differs slightly from the results of Bem (1974), who found that measuring gender as two separate scales led participants to treat gender as less binary. Moreover, where she found that masculinity and femininity were largely unrelated, we found that ratings of woman and man were strongly correlated. This is probably accounted for by the differences in outcome measures in Bem (1974) and in our study. Bem (1974) measured gender as a psychological trait in the self, whereas we measured gender as a judgment of the gender identity of others. The latter outcome is not only determined by the response options, but also by the physical features of the faces. In other words, judging the faces of others is a different task from judging one’s own characteristics, and one of the primary differences is the increase in external stimuli and influences.

The finding from Study 2 that participants use non-binary response options is consistent with the work of Saperstein and Westbrook (2021) and Lindqvist et al. (2020), which has shown that including flexible response options allows participants to better express themselves. A recommendation from that literature is that open text boxes afford participants the greatest flexibility in their responses. In our study that flexibility was rarely used when the response options consisted of a free text. This likely reflects the difference between transgender and gender-diverse participants categorizing their own gender and cisgender participants categorizing others.

A probable explanation for the difference between free text and multiple categories in Study 2 is that the multiple categories served as a visual reminder of non-binary identity. Researchers interested in the categorization of non-binary identity should be aware that these may not spring to mind unless participants are explicitly reminded of them.

Neither free text nor multiple categories influenced the categorizations of women and men. This suggests that such inclusive response options can be suitable for investigating the categorization of women and men without skewing the results or introducing noise. This is a positive finding for researchers who are primarily interested in such categorizations but do not want to contribute to the marginalization of trans and non-binary individuals.

Overall, we recommend researchers include non-binary response options in gender categorization studies. Multiple dimensions, free text, and multiple categories and continua are all viable alternatives. If the primary research question is to investigate non-binary categorization, then multiple categories are most suitable. However, if the goal is to measure the categorization of women and men, free text or multiple categories may be equally suitable.

## Limitations and future directions

One limitation of this study is the sample size. The *N*s in each condition are below many of the conventional recommendations in social psychology. However, these recommendations are typically made based on the assumption of a single trial per participant. In contrast, each participant completed 126 trials in our experiment. This allows for precise detailing of the within-participant processes. As such, the present study resembles psychophysical experiments, which also feature few participants carrying out many trials. Power is often portrayed as a function of sample size, and this is true, the number of trials is also a factor in power (Judd et al., 2017). Indeed, the overall analyses included more than 8000 data points in each experiment, and the final estimates were measured with a high degree of precision. That said, we note that the generalizability of the experiment is somewhat reduced.

Another limitation of this study is that it does not account for the influence of markers of gender other than faces. Such markers include hair, clothes, and makeup and Transgender and gender diverse often use such markers to signal their gender to others. Moreover, the faces used here were not “realistic” in that they did not realistically depict gender diversity as it is often displayed in the real world. In that sense, it is possible that we underestimate the rates of people responding with one of the options beyond the binary.

## Conclusion

In two studies, we tested how different response alternatives affected gender categorizations. In Study 1, participants responded categorically to the faces, both when rating gender using one-dimensional and two-dimensional scales. This suggests that participants generally had a binary conception of gender, which was not influenced by response options. In Study 2, participants were more likely to categorize faces beyond the binary when using multiple categories, including non-binary and I don’t know than when using a free text option. In comparison to self-identification questions, where open-ended responses are seen as the most inclusive alternative (Lindqvist et al., 2020), the categorization of others benefits from response options that explicitly remind participants that not all people identify as women or men.

# References

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

Bem, S. L. (1974). The measurment of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42*(2), 155.

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>

Butler, J. (1999). *Gender trouble: feminism and the subversion of identity*. Routledge.

Campanella, S., Chrysochoos, A., & Bruyer, R. (2001). Categorical perception of facial gender information: Behavioural evidence and the face-space metaphor. *Visual Cognition*, *8*(2), 237–262. <https://doi.org/10.1080/13506280042000072>

Campanella, S., Hanoteau, C., Seron, X., Joassin, F., & Bruyer, R. (2003). Categorical perception of unfamiliar facial identities, the face-space metaphor, and the morphing technique. *Visual Cognition*, *10*(2), 129–156. <https://doi.org/10.1080/713756676>

DeBruine, L. (2018). WebMorph. In *WebMorph*. https://webmorph.org/.

DeBruine, L., & Jones, B. C. (2017). Face Research Lab London Set. *Figshare*. <https://doi.org/10.6084/m9.figshare.5047666>

Habibi, R., & Khurana, B. (2012). Spontaneous Gender Categorization in Masking and Priming Studies: Key for Distinguishing Jane from John Doe but Not Madonna from Sinatra. *PLoS ONE*, *7*(2), e32377. <https://doi.org/10.1371/journal.pone.0032377>

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*. <https://doi.org/10.1037/amp0000307>

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with More Than One Random Factor: Designs, Analytic Models, and Statistical Power. *Annual Review of Psychology*, *68*(1), 601–625. <https://doi.org/10.1146/annurev-psych-122414-033702>

Jung, K. H., White, K. R. G., & Powanda, S. J. (2019). Automaticity of gender categorization: A test of the efficiency feature. *Social Cognition*, *37*(2), 122–144. <https://doi.org/10.1521/soco.2019.37.2.122>

Levitt, H. M., & Ippolito, M. R. (2014). Being transgender: The experience of transgender identity development. *Journal of Homosexuality*, *61*(12), 1727–1758. <https://doi.org/10.1080/00918369.2014.951262>

Liben, L. S., & Bigler, R. S. (2017). Understanding and Undermining the Development of Gender Dichotomies: The Legacy of Sandra Lipsitz Bem. *Sex Roles*, *76*(9-10), 544–555. <https://doi.org/10.1007/s11199-015-0519-4>

Lindqvist, A., Sendén, M. G., & Renström, E. A. (2020). What is gender, anyway: A review of the options for operationalising gender. *Psychology & Sexuality*, 1–13. <https://doi.org/10.1080/19419899.2020.1729844>

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. <https://doi.org/10.3758/s13428-014-0532-5>

Monro, S. (2019). Non-binary and genderqueer: An overview of the field. *International Journal of Transgenderism*, *20*(2-3), 126–131. <https://doi.org/10.1080/15532739.2018.1538841>

Morgenroth, T., & Ryan, M. K. (2021). The effects of gender trouble: An integrative theoretical framework of the perpetuation and disruption of the gender/sex binary. *Perspectives on Psychological Science*, *16*(6), 1113–1142. <https://doi.org/10.1177/1745691620902442>

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

Richards, C., Bouman, W. P., Seal, L., Barker, M. J., Nieder, T. O., & T’Sjoen, G. (2016a). Non-binary or genderqueer genders. *Int Rev Psychiatry .*, *28(1)*, 95–102.

Richards, C., Bouman, W. P., Seal, L., Barker, M. J., Nieder, T. O., & T’Sjoen, G. (2016b). Non-binary or genderqueer genders. *International Review of Psychiatry*, *28*(1), 95–102. <https://doi.org/10.3109/09540261.2015.1106446>

Saperstein, A., & Westbrook, L. (2021). Categorical and gradational: Alternative survey measures of sex and gender. *European Journal of Politics and Gender*, *4*(1), 11–30. <https://doi.org/10.1332/251510820X15995647280686>

Simanova, I., Francken, J. C., de Lange, F. P., & Bekkering, H. (2016). Linguistic priors shape categorical perception. *Language, Cognition and Neuroscience*, *31*(1), 159–165. <https://doi.org/10.1080/23273798.2015.1072638>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. <https://doi.org/10.21105/joss.01686>

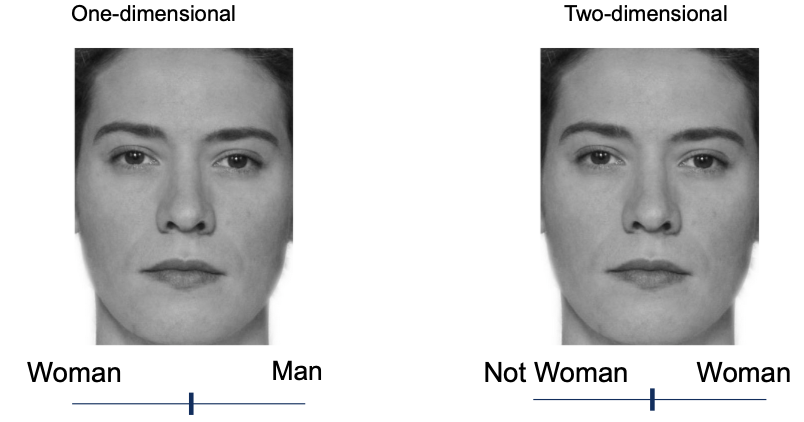
Figure 1

Example of a seven-step morphing spectrum



Figure 2

Sample trial from each of the three conditions



Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':  
  
 combine

Figure 3

Participant level and mean ratings of faces in One-dimensiona and two-dimensional conditions

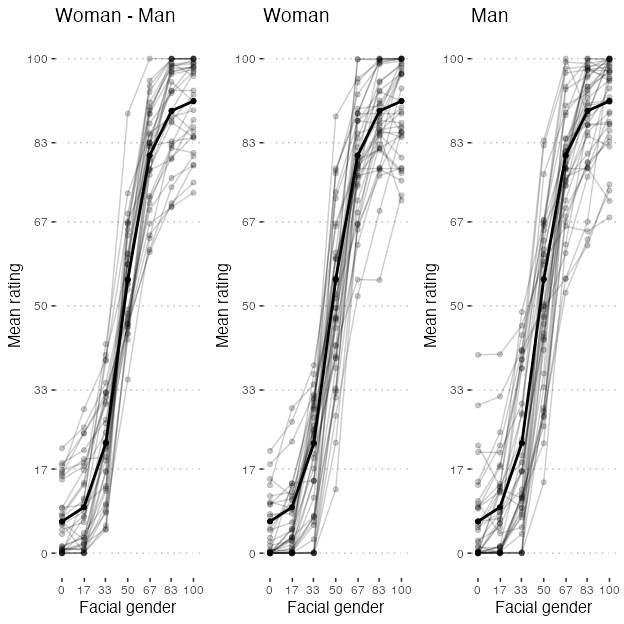


Figure 4

Sample trial from each of the three conditions

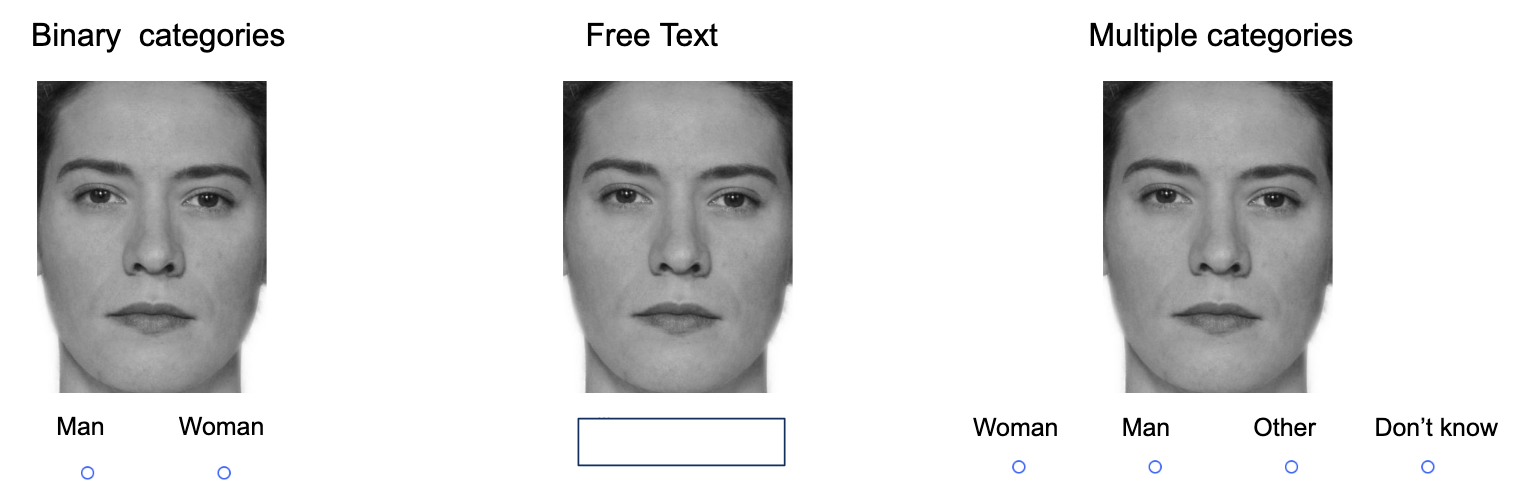


Figure 5

Gender Categorizations by Participants

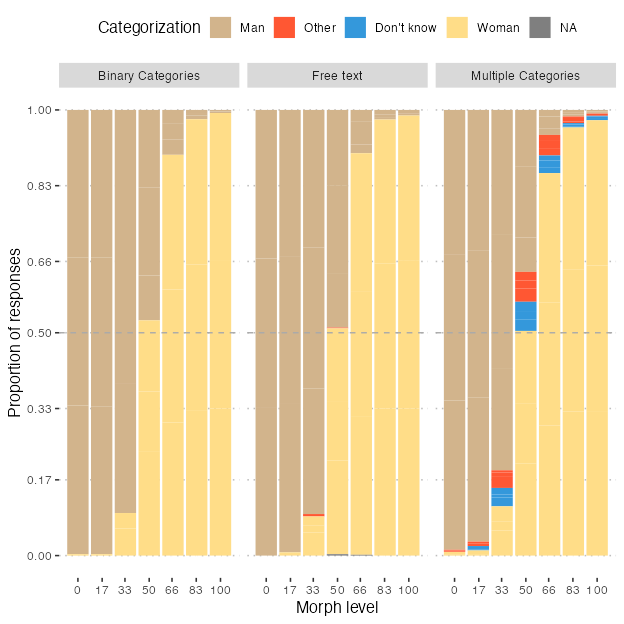


Figure 6

Responses of other and I don't know across the multiple categories and free text condiitons

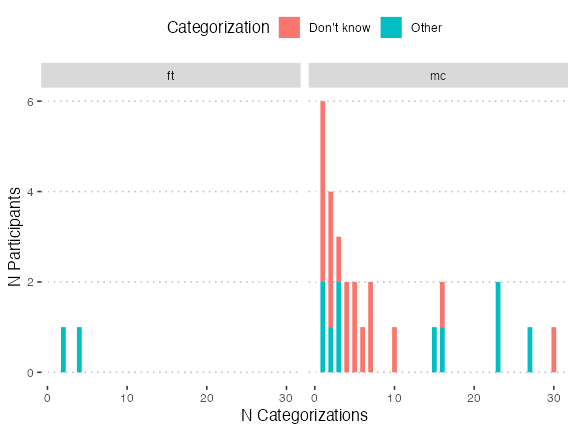


Figure 7

Alternative version of the previous figures

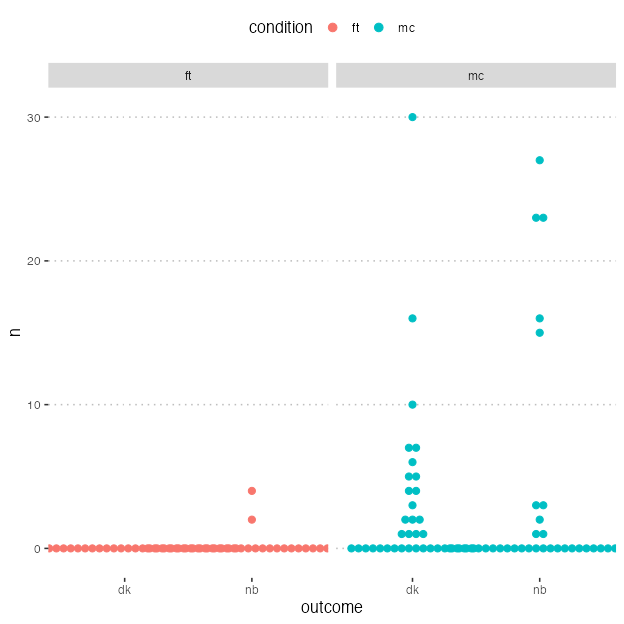


Figure 8

Participant Proportions for Categorizing Faces as Women Across Three Conditions

