Supplementary Material for *Flexibility in continuous judgements of gender/sex and race*

Table of Contents

Additional Information about Studies 1a and 1b

Study 1a and 1b Detailed Method

In both Studies 1a and 1b, participants saw a black line (i.e., continuum) with labels on either end. On each trial, participants saw a stimulus (representing the domain of gender, color, or number; see examples in Figure 1 in main manuscript and see below for detailed information on stimuli for all studies) and had to click the location along the continuum where they thought that item was best represented in relation to the labels at either end. Additionally, participants were instructed to sort as quickly as possible without losing accuracy, and were given a maximum of 4000ms to respond to each stimulus. For gender trials (see manuscript Figure 1, Panel B), the target stimuli were morphed faces that were Asian, Black, or White, and the labels were the words 'female' and 'male'. For color trials, the target stimuli were colors and the labels were the end point colors from which the target stimulus was created. Finally, for number trials, the target stimuli were numbers, and the labels were the end point numbers from the set to which the target stimulus belonged.

Since there were three times as many gender stimuli as color or number stimuli in studies 1a and 1b, we split them up across 3 blocks of 54 trials each and interspersed the color and number blocks (which were counterbalanced across participants and also contained 54 trials each) between them. Each of the 270 stimuli was sorted exactly once, and all stimuli were presented in a random order within each category domain. Additionally, all participants saw an equal number of Asian, Black, and White faces. In both studies, labels were stationary (e.g. "female" was always on the left), though this varied in later studies. In Study 1a, gender stimuli were randomized across all gender blocks, but in Study 1b, gender stimuli were grouped based on the pair of faces used to generate them (e.g. all 9 faces generated by a single male/female pair were always run in the same block) to ensure our observed results were not design-dependent.

In both studies, data were collected via Inquisit Web (Millisecond Software, 2016). The resulting data files included two variables of interest; latency (i.e. how long it took for participants to sort the target stimulus; all latency results are reported below), and subjective judgment (i.e. the place on the bar that people clicked to sort each stimulus relative to the end points). Tasks were programmed so that clicking the leftmost edge of the line was recorded as 0, and clicking the rightmost edge of the line was recorded as 100, and anywhere in between was recorded as a value between 0-100 proportional to the place along the line that was clicked. In all cases, we computed participants' accuracy based on their judgment error for each trial (i.e. the absolute value of the difference between the subjective judgment and objective location). For instance, for an 80/20 morph, the objective location is 80% of the way across the bar so if a participant's subjective judgment was at 75%, the judgment error would be 5).

Studies 1a and 1b Stimuli

Gender stimuli for Studies 1a and 1b were created by randomly pairing the 12 most masculine male faces and 12 most feminine female faces in each of 3 races (Asian, Black, and White) from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015). We cropped each selected face into an oval containing only the face (i.e. excluded all background color, neck, ears, hair, etc) and uploaded it onto the WebMorph interface (DeBruine, 2018) and delineated each face using the 'auto-delineate' function which automatically identified 93 anchor points on each face. The resulting 36 male-female, monoracial face

pairs served as the end points, which were systematically morphed using WebMorph's "transform" function (DeBruine, 2018) at a 10% gradient to create 9 distinct morphs from each of the 36 pairs of faces (e.g., 10% of the female face and 90% of the male, etc; see Figure 1 in manuscript for an example set of face gender stimuli)[1]. At test, participants in Study 1a sorted all morphs from half the face pairs of each race (162 faces total), and morphs from the other half of the face pairs were given to a different set of participants in Study 1b (see https://osf.io/u69yj/?view_only=c9374011248b42fcb1c257242a775b23 for full sets of face stimuli for both studies). This meant that Study 1b was a replication with unique stimuli (in line for calls for stimulus sampling; Wells & Windschitl, 1999).

Color stimuli for Studies 1a and 1b were created by combining primary colors (e.g. red, green, blue) and secondary colors (e.g. yellow, cyan, magenta) with one another to form 6 distinct pairs of end points. Each of the 6 colors was represented twice to create 6 total continua, and we purposefully avoided pairing complementary colors due to their propensity to cancel eachother out (or lose hue) and produce a grayscale color when mixed. Then, each pair was systematically morphed from at a 10% gradient, resulting in 9 intermediary colors that were equidistant from one another in RGB color space for each of the 6 pairs (see Figure 1 for an example set of color stimuli). At test, participants sorted all 54 color stimuli in both Studies 1a and 1b (see https://osf.io/u69yj/?view_only=c9374011248b42fcb1c257242a775b23 for a full set of color stimuli).

Number stimuli for Studies 1a and 1b were created by using a multi-step process. First, we used random.org to generate 6 distinct multiples of 10 between 0-100 to serve as lower values. Then we generated 6 numbers between 1 and 100, which served as morphing intervals. Then, for each lower value, we systematically added the morph interval 10 times to generate the various gradations that cumulatively mirrored the ones created for the face and color stimuli. This resulted in 9 intermediary values that were equidistant from one another for each of the 6 number pairs (see Figure 1 for an example set of number stimuli. At test, participants sorted all 54 number stimuli in both Studies 1a and 1b (see https://osf.io/u69yj/?view_only=c9374011248b42fcb1c257242a775b23 for a full set of number stimuli).

Study 1a Analyses

Study 1a pre-registered main analyses[2]

"We will conduct a one-way anova comparing mean latency between color, [gender], and number trials, and will follow-up with posthoc Tukey tests if initial anova is significant."

A one-way anova comparing mean latency between color, gender, and number trials revealed significant differences in speed across the three domains, $F(2,25764) = 1573.00$, $p < .001$, and planned post-hoc Tukey Tests revealed that participants were faster at judging gender than at judging number or color, and also significantly faster at judging color than number (all p's $< .001$)

"We will conduct a one-way anova comparing mean [accuracy] between color, [gender], and number trials, and will follow-up with posthoc Tukey tests if initial anova is significant."

[1] Note that we did not include the original faces used to make the morphs as stimuli during the task

[2] Because the specific terminology we used shifted over time, we have standardized this document to align with the equivalent terminology used in the main manuscript for clarity and consistency. Any words that were changed from the original pre-registration will appear in [brackets].

A one-way anova comparing color, gender, and number trials revealed significant differences in accuracy across the three domains, $F(2,25764) = 459.40$, $p < .001$, and planned post-hoc Tukey Tests revealed that participants were less accurate at judging gender than at judging number or color, and also significantly less accurate at judging color than number (all p's < .001)

"We will separate the data by domain (color, [gender], and number) and then generate various models (e.g. linear, one-cycle, or two-cycle), apply each model to each domain in order to determine which model best explains the relation between [subjective] judgment and [objective location]."

For each domain (color, gender, and number), we evaluated the fit of a linear model, as well as the standard one- and two-cycle versions of the proportional power model used in Slusser, Santiago, and Barth, 2013. For both the gender and color data, the linear model fit the best (R^2 = .483 for gender data, R^2 = .610 for color data), followed by the two-cycle model (R^2 = .479 for gender data, R^2 = .609 for color data) and one-cycle model (R^2 = .469 for gender data, R^2 = .607 for color data). However, these fits are close in magnitude and have low R^2 values, indicating low explanatory power for these data. For number data, the linear model (R^2 = .623) was also the best fit, followed by the one-cycle model (R^2 = .608) and then the two-cycle model (R^2 = .571). Additionally, the value of the β = parameter for the one-cycle models was highest for color data (β = .903), followed by gender data (β = .835), and number data (β = .721), the most bias in estimates of colors, then gender, and least bias in estimates of numbers. This pattern was partly mirrored in the two-cycle models, which found the highest bias in color data (β = .769), followed by number data ((β = .774), and then gender data (β = .542).

Study 1a pre-registered additional analyses

"We will run 2 one-way anovas, one comparing mean [accuracy] and the other comparing mean latency between black, white, and asian trials. As above, we will follow-up with posthoc Tukey tests for any initial anova that is significant."

Using a one-way anova comparing Asian, Black, and White faces, we found a significant difference in accuracy across faces of different races, $F(2,15424) = 41.00$, $p < .001$. Additionally, planned post-hoc Tukey Tests revealed that participants were significantly more accurate when judging White faces than when judging Black or Asian faces in both studies (all p's <.001), and also significantly more accurate when judging Asian faces than Black faces (p = .014). Additionally, a one-way anova comparing mean latency between Asian, Black, and White faces found no difference in the length of time it took participants to judge faces of different races, $F(2,15424) = 2.09$, p = .124.

"We will also assess whether participants are more accurate (i.e., show smaller difference scores) for faces of their own race by re-coding [gender] stimuli as "participant's race" or "not participant's race" and computing a paired t-test on the difference scores."

Using a two-tailed paired t-test, we found that monoracial participants in both studies were more accurate in their subjective judgments of faces of their own race than faces of other races, $t(81) = 2.07$, p = 0.041.

Study 1a exploratory analyses

In order to investigate whether or not the accuracy of participants judgements varied by race, we ran a multilevel regression model on data from [gender] trials predicting error from race of the face,

location (distance from the nearest end-point), and their interaction. As in the main papers we also included a random intercept for participant ID, as well as a random slope within-participant for location. Replicating analyses for the main paper, we found that there was overall greater error for more intermediary stimuli ($b = 0.204$, $t(122) = 5.679$, $p < .001$). Additionally we found that that this effect was exacerbated for racial minorities in comparison to White faces, resulting in greater perceptual error for Black ($b = 0.131$, $t(15{,}030) = 6.810$, $p < .001$) and Asian ($b = 0.087$, $t(15{,}030) = 4.503$, $p < .001$) faces with intermediary gender/sex, relative to that of White faces with intermediary gender/sex.

Study 1b Analyses

Study 1b pre-registered main analyses

"We will conduct overall pearson correlation between [subjective judgment] and [objective] location for each subject, report both the average correlation across subjects as well as the number of individuals showing a positive vs. negative correlation, and then do a chi square test to see if this distribution of correlations is significantly different from chance (e.g. half positive, half negative)."

We found that all 99 participants had a positive correlation between subjective judgment and objective location, with a mean Pearson correlation of $r = .788$ across all participants. This distribution of responses is significantly different from the 50/50 distribution that would be expected by chance ($X^2(1, n = 99) = 99$, $p < .001$).

"*We will re-code [gender] stimuli as "participant's race" or "not participant's race" and will compute an average accuracy score (see above) for these categories. We will then compute a paired t-test comparing these accuracy scores. Any multi-racial participants will be excluded from this round of analyses."*

Replicating Study 1a, a two-tailed paired t-test again revealed an own-race face effect, such that participants were more accurate at judging faces of their own race than faces of other races ($t(91) = 3.05$, $p = .003$).

"We will run two linear regressions predicting latency from [objective] location. The first regression will isolate responses to stimuli with [objective] locations on the left side (from 0 - 50) where a significant upward slope indicates slower categorization of more androgynous faces. The second regression will isolate responses to stimuli with [objective] locations on the right side (50-100) where a significant downward slope indicates faster categorization of more less androgynous faces."

The results of a linear regression predicting latency from objective location on the left half of the distribution (from 0 - 50) revealed that participants took longer to categorize faces as they became more androgynous, $b = 4.00$, $t(1) = 8.52$, $p < .001$. A second linear regression predicting latency from objective location on the right half of the distribution (from 50 - 100) revealed that participants took less time to categorize faces as they became less androgynous, $b = -5.05$, $t(1) = -11.16$, $p < .001$. In other words, both sides showed the same pattern such that more androgynous faces were slower to place along the continuum.

Study 1b pre-registered additional analyses

"*We will calculate a correlation between participants' mean accuracy for [gender] data and GSDB score, Transphobia score, and the difference score on the thermometer ratings (cis minus trans) respectively."*

There were no significant correlations between participants' accuracy at judging gender and levels of transphobia ($r(98) = 0.19$, $p = .059$), gender/sex essentialism ($r(98) = -0.08$, $p = .415$), or feelings towards cisgender and transgender people ($r(98) = -0.09$, $p = .398$).

"*We will conduct a one-way anova comparing mean latency between color, [gender], and number trials, and will follow-up with posthoc Tukey tests if initial anova is significant."*

The results of one-way anova comparing mean latency between color, gender, and number trials replicated Study 1a, revealing significant differences in speed across the three domains, $F(2,26272) = 1229.00$, $p < .001$, and planned post-hoc Tukey Tests revealed that participants were slower at judging number than gender and color (all p's < .001). However, unlike in Study 1a, there was no difference in latency when judging color compared to gender.

"*We will conduct a one-way anova comparing mean accuracy between color, [gender], and number trials, and will follow-up with posthoc Tukey tests if initial anova is significant."*

The results of one-way anova comparing mean accuracy between color, gender, and number trials fully replicated Study 1a, revealing significant differences in speed across the three domains, $F(2,26272) = 325.40$, $p < .001$, and planned post-hoc Tukey Tests again revealed that participants were less accurate at judging gender than at judging number or color, and also significantly less accurate at judging color than number (all p's < .001).

Study 1b exploratory analyses

Like in Study 1a, we investigated whether or not the accuracy of participants judgements varied by race by running a multilevel regression model on data from [gender] trials predicting error from race of the face, location (distance from the nearest end-point), and their interaction. Again, we also included a random intercept for participant ID, as well as a random slope within-participant for location. Replicating study 1a, we found that there was overall greater error for more intermediary stimuli ($b = .260$, $t(128) = 5.679$, $p < .001$). This time, we found that that this effect was exacerbated for Black faces in comparison to White faces, resulting in greater perceptual error for Black ($b = .056$, $t(15,500) = 3.242$, $p = .001$) faces with intermediary gender/sex, relative to that of White faces with intermediary gender/sex. Unlike Study 1, we found no differences in judgements of Asian faces in comparison to White faces.

Studies 1a & 1b Brief Summary of Findings

Across multiple participant samples, stimuli sets, and study designs, we provide evidence that when asked to sort faces (ranging in gender phenotype), colors, and numbers along a line with labeled endpoints, results demonstrate robust and distinct patterns of inaccurate subjective judgments of gender. We find that despite positive correlations between subjective judgment and objective location in all three domains, people are consistently more inaccurate when judging gender than when judging color and number. However, we find no association between task accuracy and social correlates. Additionally, we twice find evidence of differences in perceptions based on face race reminiscent of previous work on the own-race face effect (see MacLin & Malpass, 2001; Golby, Gabrieli, Chiao, & Eberhardt, 2001). Finally, we find preliminary evidence (using data from Study 1b) that people may be slower to categorize faces that are more androgynous (presumably because their phenotype is more ambiguous in comparison to those that are either clearly masculine or feminine). In Study 2, we sought to replicate and extend the patterns found in perception of gender to race.

Additional Information about Study 2

Study 2 Detailed Method

The procedure for Study 2 was very similar to the set-up used in Studies 1a and 1b. In this iteration of the task, participants sorted 2 blocks of race stimuli (which contained equal numbers of faces that were male and female) and 2 blocks of gender stimuli (which contained equal numbers of faces that were Asian, Black, or White) in random order each containing 54 trials. Like in Study 1b, stimuli were grouped together based on the base pair used to create them. Additionally, we counterbalanced the location of labels (which were male and female for gender trials and Asian, Black, or White for race trials) so that participants did not associate a label with one particular side of the continuum (e.g, female faces were not always on the left). As before, data were collected using Inquisit Web (Millisecond Software, 2016) and the task was programmed to collect the same variables of interest measured in Studies 1a and 1b. Also like in prior studies, participants were asked to fill out a short survey following completion of the task which included measures of social dominance orientation (Pratto et al., 1994), need for closure (Roets & Van Hiel, 2011; Webster & Kruglanski, 1994), essentialism (Bastian & Haslam, 2006), and basic demographic information.

Study 2 Stimuli

Gender stimuli were created using a nearly identical method to that described in Study 1a. The only substantial differences were that we morphed fewer faces (n = 24), and used base faces from the London Face Database (DeBruine & Jones, Benedict, 2017) to enhance stimulus sampling and generalize our findings beyond faces in the Chicago Face Database (Wells & Windschitl, 1999). The resulting set of gender stimuli were cropped and anchored faces on WebMorph using the same procedure as in Study 1. The resulting 4 same-race face pairs from each of 3 racial groups (Black, East Asian, and White) were morphed using WebMorph at a 10% gradient for a total of 108 faces that participants sorted at test.
Race stimuli were created by selecting, cropping, and anchoring 4 faces male and 4 female faces that were most consistently rated as Asian, Black, and White (for a total of 12 male faces and 12 female faces, evenly spread across 3 races) from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015). Then, we semi-randomly paired faces across race (i.e. Asian-White, Black-Asian, and White-Black) to end up with 2 male and 2 female pairs for each cross-race combination. The resulting 12 face pairs served as the end points, which were systematically morphed using WebMorph at a 10% gradient to create 9 distinct morphs (see Figure 1 in manuscript for an example set of race stimuli). At test, participants sorted all morphs (but not original faces) for a total of 108 faces (see https://osf.io/u69yj/?view_only=c9374011248b42fcb1c257242a775b23 for full set of race stimuli).

Study 2 Analyses

Study 2 pre-registered main analyses

"We will first split the data by condition (gender vs. race phenotype trials) and conduct two overall pearson correlations between [subjective judgment] and [objective] location for each subject, report both the average correlation across subjects as well as the number of individuals showing a positive vs.

negative correlation, and then do a chi square test to see if this distribution of correlations is significantly different from chance (e.g. half positive, half negative).”

We found that all 148 participants had a positive correlation between subjective judgment and objective location for race with a mean Pearson correlation of $r = .814$. This distribution of responses is significantly different from the 50/50 distribution that would be expected by chance ($X^2(1, n = 148) = 148$, $p < .001$). Of participants with eligible data from gender trials, all 147 had a positive correlation between subjective judgment and objective location, with a mean Pearson correlation of $r = .722$. This distribution of responses is significantly different from the 50/50 distribution that would be expected by chance ($X^2(1, n = 147) = 147$, $p < .001$).

“We will re-code gender phenotype stimuli as “participant’s race” or “not participant’s race” and will compute an average accuracy score (see above) for these categories. We will then compute a paired t-test comparing these accuracy scores. Any multi-racial participants will be excluded from this analysis.”

Unlike the two previous studies, we did not replicate the own-race face effect within the gender task, meaning that participants in Study 2 were no more accurate at judging faces of their own race than faces of other races; a two-tailed paired t-test revealed $t(132) = 1.208$, $p = .229$.

“*We will split the data by condition (gender vs. race phenotype trials) and then create a new ambiguity score for each trial based on how far away the [objective] location is away from its closest anchor, 0 or 100. Then we will run two linear regressions predicting accuracy and latency from ambiguity score for race phenotype trials and two for gender phenotype trials.”*

The results of a linear regressions predicting latency from ambiguity revealed that participants took longer to categorize faces as they became more phenotypically ambiguous ($b = 5.22$, $t(1) = 14.44$, $p < .001$ for race trials and ($b = 4.30$, $t(1) = 11.38$, $p < .001$ for gender trials). A second pair of linear regression predicting accuracy from ambiguity revealed that participants were less accurate at categorizing faces as they became more phenotypically ambiguous ($b = .22$, $t(1) = 31.74$, $p < .001$ for race trials and $b = .25$, $t(1) = 28.66$, $p < .001$ for gender trials).

Study 2 pre-registered additional analyses

“We will calculate a correlation between participants' mean task accuracy and social dominance orientation, need for closure, and psychological essentialism, respectively.”

There were no significant correlations between participants’ mean task accuracy and social dominance orientation ($r(131) = -0.04$, $p = .628$), need for closure ($r(131) = 0.14$, $p = .118$), or psychological essentialism ($r(131) = 0.06$, $p = .516$).

Study 2 exploratory analyses

Like in Study 1a and 1b, we investigated whether or not the accuracy of participants judgements varied by race by running a multilevel regression model on data from [gender] trials predicting error from race of the face, location (distance from the nearest end-point), and their interaction. As before, we also included a random intercept for participant ID, as well as a random slope within-participant for location. Replicating study 1a, we found that there was overall greater error for more intermediary stimuli ($b = .224$, $t(134) = 6.79$, $p < .001$). Like before, we found that that this effect was exacerbated for Black faces in comparison to White faces, resulting in greater perceptual error for Black ($b = .047$, $t(15{,}360) = 2.475$,

p = .013) faces with intermediary gender/sex, relative to that of White faces with intermediary gender/sex. Unlike Study 1a and 1b, we also found that in comparison to White faces, people exhibited overall greater perceptual error categorizing Black faces (b = 1.208, t(15,360) = 2.06, p = .039), and marginally greater perceptual error categorizing Asian faces (b = 1.16, t(15,370) = 1.98, p = .047).

Study 2 Brief Summary of Findings

This study tested whether select results from Studies 1a and 1b would replicate and generalize to an iteration of our task that varied along the domains of race and gender phenotype (as opposed to number and color). First, we replicated the correlation between subjective judgment and objective location observed in Study 1b, finding a strong, positive correlation between subjective and objective location in both domains. Additionally, we sought to replicate the own-race face bias observed in Studies 1a and 1b, but found instead that people were no more accurate at judging faces of their own race than faces of other races. Despite lack of replication, this result does not rule out the possibility of a true effect and previous research has demonstrated a high probability of obtaining mixed results (including failed replications) even when an effect is true (Lakens & Etz, 2017). Finally, building off our findings from Study 1b, we found that participants were both slower and less accurate judging more intermediary faces (i.e. those that were closer to the middle of the distribution) but again, no relationship between task accuracy and (a different set of) survey measures.

Additional Information about Study 3

Study 3 Detailed Method

In Study 3, participants were first asked to self-select into one of three participant groups (Gender Diverse, Cisgender-LGBPQ+, or Cisgender Heterosexual), and then proceeded to complete a sorting task with a similar set-up to those used in Studies 1a, 1b, and 2. In this iteration of the task, participants sorted 2 blocks of stimuli by gender/sex, both of which contained equal numbers of faces that were Asian, Black, or White. Like Studies 1b and 2, each block contained 54 trials that were grouped together based on the base pair used to create them and we counterbalanced the location of labels so that participants did not associate a label with one particular side of the continuum. The key difference between the two blocks of stimuli was that we manipulated the amount of time participants had to sort each face. In one block, participants were given 4000ms per face (dubbed the “slow” condition and is equivalent to the amount of time per trial in Studies 1a, 1b, and 2) and in the other block participants were given only 2000ms per trial (dubbed the “fast” condition; also see Study 3 Pilot Study below). All participants completed one block in each condition in a random order (counterbalanced across subjects). As before, data were collected using Inquisit Web (Millisecond Software, 2016) and the task was programmed to collect the same variables of interest measured in Studies 1a, 1b, and 2. Also like in prior studies, participants were asked to fill out a short survey following completion of the task which included measures of gender identity reflection and rumination (Bauerband & Galupo, 2014), theoretical awareness of genderqueer identity (McGuire et al., 2019), and basic demographic information.

Study 3 Pilot Study

In order to determine if participants' sorting patterns change when sorting faces in a more rushed way, we manipulated the amount of time participants had to sort each face. To determine the time allotted per trial in the rushed condition, we conducted a 20 person pilot study (not pre-registered) on Amazon's Mechanical Turk in Summer 2021 comparing participants' responses when given 2000ms vs. 4000ms to ensure we maximized the number of trials participants responded to while still creating the feeling of being "rushed".

Study 3 Stimuli

Similar to Studies 1a, 1b, and 2, stimuli for Study 3 was created by randomly cropping, anchoring, and pairing the 4 of the most phenotypically masculine and 4 of the most phenotypically feminine faces within Asian, Black and White races from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015), and morphing them in WebMorph at a 10% gradient (see Figure 1 in main manuscript for example). This resulted in 108 faces (9 distinct morphs from 12 face pairs), which participants sorted across 2 blocks in Study 3.

Study 3 Demographics Table (including cis-queer participants)

In Study 3 we measured gender/sex and identity and sexual orientation at two points during the study. First, participants were asked to self-select into one of the three participant groups (Gender Diverse, Cisgender-LGBPQ+, or Cisgender Heterosexual) at the beginning of the study. Second participants were asked to write in their current gender/sex identity and sexual orientation via free-response boxes as part of the demographics section at the end of the survey. Participants' answers to the demographic questions about gender/sex identity and sexual orientation did not always align with group self-selection (e.g. a participant who self-selected into the "Cisgender Heterosexual" group but listed their sexual orientation as bisexual). Notably, analyses excluding these participants revealed no substantial changes in results, so we chose to retain participants' initial self-selected category. Due to the open-ended format of our gender/sex and sexual orientation questions, some participants self-reported multiple identity labels. In these cases, we counted participants in all the categories they listed (e.g. a participant who self-reported their gender as agender and nonbinary would be counted as both the "agender" and "nonbinary").

Table 1. Demographic information for participants in Study 3

Participant Group	Gender Diverse (*n* = 194)	Cisgender-LGBPQ+ (*n* = 196)	Cisgender Heterosexual (*n* = 197)
Age			

Mean (in years)	24.8	24.3	25.44
SD (in years)	5.5	5.8	7.8
Race			
Asian	7	5	7
Black	5	8	24
Latinx	17	61	36
White	142	105	116
Multiracial	20	13	14
Other	3	4	0
Gender/Sex Identity			
Agender	6%	1%	0.5%
Bigender	2%	0.5%	0.5%
Demigender	1%	0.5%	0%
Female / Woman	9%	48%	46%
Genderfluid	8%	0.5%	0%
Gender Non-Conforming	3%	1%	0%
Genderqueer	7%	0.5%	0%
Male / Man	11%	42%	51%
Non-Binary	44%	2%	0%
Trans Woman/Femme	5%	0%	0%
Trans Man/Masc	10%	0%	0%
Other	3%	4%	2%
Self-Reported Multiple Gender Identities	9%	2%	0%
Sexual Orientation			

Asexual / Aromantic	13%	0.5%	0.5%
Bisexual	34%	59%	2%
Demisexual	1%	3%	1%
Gay	4%	11%	0%
Heterosexual / Straight	6%	2%	86%
Homosexual	4%	9%	0%
Lesbian	9%	4%	0%
Pansexual	19%	6%	0%
Polyamorous / Polysexual	1%	1%	0%
Queer	12%	2%	0%
Other	3%	3%	10%
Self-Reported Multiple Sexualities	8%	0.5%	0.5%

Note: Percentages reflect participants' ability to self-report multiple gender identity/sexual orientation labels.

Study 3 Analyses

Study 3 pre-registered main analyses

"We will conduct a multiple linear regression on gender diverse and cisgender-heterosexual participant data, predicting accuracy from the following predictors of interest: androgyny level, response window, and participant group. To do this, we will use a fully-specified mixed model (listed below) which will include fixed effects of androgyny, response window, participant group and their 2 and 3-way interactions. The model will also include a random intercept for subject id, as well as random slopes within-participant for androgyny level, response window, and their interaction. We will also include a random intercept for stimulus face-pair."

*Model: accuracy ~ androgyny_level*response_window*participant_group + (androgyny_level*response_window | subjectID) + (1|face_pair)*

As reported in the manuscript, the results from a multiple linear regression revealed that participants were overall less accurate at categorizing faces that were more androgynous, (b = .148, t(450.7) = 12.900, p < .001), but that this effect was stronger for cisgender heterosexual participants than for gender diverse participants (that is, gender diverse participants were less bad at categorizing more androgynous faces than cisgender heterosexual participants; b = -.081, t(452.2) = - 4.968, p < .001).

"Regardless of the outcome, we will also re-run the above model (omitting the effects of androgyny level) on the following subsets of participant data:

- *Gender diverse and cis-het participant data when androgyny level = 10*
- *Gender diverse and cis-het participant data when androgyny level = 20*

- *Gender diverse and cis-het participant data when androgyny level = 30*
- *Gender diverse and cis-het participant data when androgyny level = 40*
- *Gender diverse and cis-het participant data when androgyny level = 50"*

While there were no differences between participant groups when sorting less androgynous faces (Androgyny Levels 10 and 20), gender diverse participants were significantly more accurate at categorizing more androgynous faces (beginning at Androgyny Level 30 and above, also see Table 2 below).

Table 2. *Main effect of accuracy across participant groups (cisgender heterosexual vs. gender diverse) at each level of androgyny*

Androgyny Level	*Estimate*	Std. Error	*df*	*t*	*p*
Androgyny Level 10	.01	.01	381.48	1.43	.152
Androgyny Level 20	-.007	.007	371.03	-.98	.327
Androgyny Level 30	-.04	.007	381.00	-5.07	<.001***
Androgyny Level 40	-.05	.009	382.68	-5.77	<.001***
Androgyny Level 50	-.06	.01	383.15	-5.77	<.001***

"We will run a multi-level linear regression (listed below) predicting accuracy from participant group on gender diverse and cis-het participant data when androgyny level = 50 in the Slow trials."

Model: accuracy ~ participant_group + (1|subjectID)

We find when given more time (4000ms), gender diverse participants are more accurate at cisgender heterosexual participants at categorizing the most androgynous faces (b = -.056, t(361.27) = -5.198, p < .001).

Study 3 pre-registered additional analyses

"We will run the fully-specified model listed in the Main Analyses on the full dataset but this time we will include all three participant groups (using the cis-LGBPQ+ participant group as the reference group)."

Results from a multiple linear regression including the cis-LGBPQ+ participants revealed that like before, participants were overall less accurate at categorizing faces that were more androgynous, (b = .148, t(681.7) = 13.372, p < .001), but that this effect was stronger for cis-LGBPQ+ participants than for gender diverse participants (that is, gender diverse participants were less bad at categorizing more androgynous faces than cis-LGBPQ+ participants; b = -.081, t(684.1) = - 5.15, p < .001). There were no significant differences between cis-LGBPQ+ participants and cisgender heterosexual participants.

"We will also run the fully-specified model listed in the Main Analyses section using the cis-LGBPQ+ participant group as the reference group on the following subsets of participant data:

- *Full dataset when stimulus androgyny level = 10*
- *Full dataset when stimulus androgyny level = 20*
- *Full dataset when stimulus androgyny level = 30*
- *Full dataset when stimulus androgyny level = 40*
- *Full dataset when stimulus androgyny level = 50"*

While there were no differences between participant groups when sorting least androgynous faces (Androgyny Level 10), cis-LGBPQ+ participants were significantly less accurate at categorizing more androgynous faces (beginning at Androgyny Level 20 and above when compared to cisgender heterosexual participants and beginning at Androgyny Level 30 and above when compared to gender diverse participants; also see Table 3 below).

Table 3. *Main effect of accuracy across participant groups (cisgender LGBPQ+ v. cisgender heterosexual and cisgender LGBPQ+ vs. gender diverse) at each level of androgyny*

Androgyny Level	Reference Group	Comparison Group	*Estimate*	Std. Error	*df*	*t*	*p*
Androgyny Level 10	Cisgender LGBPQ+	Cisgender Heterosexual	-.01	.01	564.96	-1.25	.214
Androgyny Level 10	Cisgender LGBPQ+	Gender Diverse	.01	.01	558.54	1.52	.130
Androgyny Level 20	Cisgender LGBPQ+	Cisgender Heterosexual	-.01	.007	549.50	-2.05	.041*
Androgyny Level 20	Cisgender LGBPQ+	Gender Diverse	-.01	.007	546.06	-1.02	.308
Androgyny Level 30	Cisgender LGBPQ+	Cisgender Heterosexual	-.02	.007	567.04	-2.68	.007**
Androgyny Level 30	Cisgender LGBPQ+	Gender Diverse	-.04	.007	556.38	-5.08	<.001***

Androgyny Level 40	Cisgender LGBPQ+	Gender Diverse	-.02	.009	572.75	-2.59	.001**
Androgyny Level 40	Cisgender LGBPQ+	Cisgender Heterosexual	-.05	.009	563.46	-5.83	<.001***
Androgyny Level 50	Cisgender LGBPQ+	Cisgender Heterosexual	-.01	.01	571.20	-1.44	.150
Androgyny Level 50	Cisgender LGBPQ+	Gender Diverse	-.05	.01	561.26	-5.91	<.001***

"We will use the same models as we did for accuracy, switching to this new dependent variable of latency (listed below below). We will first use this model only with data from cisgender-heterosexual and gender diverse participants, and then use the same model but include data from cisgender-LGBPQ participants with them as the reference group). However, we will not conduct further analysis on subsets of data defined by androgyny…unless androgyny interacts significantly with response window or participant group in the overall predictive models for latency."

*Model: latency ~ androgyny_level*response_window*participant_group + (androgyny_level*response_window | subjectID) + (1|face_pair)*

When including only cisgender-hterosexual and gender diverse participants, we find that participants are overall slower at categorizing faces when given a longer window to respond (4000 ms as opposed to 2000 ms; $b = 142.49$, $t(520.9) = 7.944$, $p < .001$). Additionally, we find this is effect is accentuated when categorizing more androgynous faces (that is, participants who were given a longer response window were especially slow to categorize more androgynous faces in comparison to participants who were given a shorter response window; $b = 82.35$, $t(35982.35) = 4.806$, $p < .001$). Due to the aforementioned interaction between androgyny level and response window, we separated the data by androgyny level and re-ran the model (omitting the effect of androgyny level) comparing across response window (long vs. short) at each level of androgyny. At every level of androgyny, we found that participants took longer to categorize faces when given more time, also see Table 4 below).

Table 4.. *Main effect of latency across response window (4000ms vs. 2000ms) at each level of androgyny*

Androgyny Level	*Estimate*	Std. Error	*df*	*t*	*p*
Androgyny Level 10	154.87	16.64	339.21	9.31	<.001***

Androgyny Level 20	152.71	16.95	344.24	9.01	<.001***
Androgyny Level 30	186.04	20.20	351.29	9.21	<.001***
Androgyny Level 40	214.82	21.49	349.78	10.00	<.001***
Androgyny Level 50	219.31	25.50	351.79	8.60	<.001***

After adding cis-LGBPQ+ participants to the model, we replicate the latency effects above. Specifically, we find that participants are overall slower at categorizing faces when given a longer window to respond ($b = 143.41$, $t(776.33) = 7.940$, $p < .001$), and that this effect is accentuated when categorizing more androgynous faces ($b = 82.80$, $t(53441.09) = 4.813$, $p < .001$). As before, when separated by androgyny level, we find that participants took longer to categorize faces when given more time at every level of androgyny, also see Table 5 below).

Table 5. *Main effect of latency across response window (4000ms vs. 2000ms) at each level of androgyny*

Androgyny Level	*Estimate*	Std. Error	*df*	*t*	*p*
Androgyny Level 10	155.23	16.74	507.38	9.27	<.001***
Androgyny Level 20	152.75	17.42	514.54	8.77	<.001***
Androgyny Level 30	185.19	20.71	528.83	8.94	<.001***
Androgyny Level 40	215.06	21.46	518.13	10.02	<.001***
Androgyny Level 50	219.96	25.03	518.50	8.79	<.001***

Notably, there were no differences in latency between gender groups in any of our models.

"We will look at overall correlations between each individual difference measure (GRRS and GQI) and each measure of task performance (accuracy and latency) by calculating the following correlations between the following items within each participant group:

- *Mean accuracy on Fast trials and GRRS score*
- *Mean accuracy on Fast trials and GQI score*
- *Mean accuracy on Slow trials and GRRS score*

- *Mean accuracy on Slow trials and GQI score*
- *Mean latency on Fast trials and GRRS score*
- *Mean latency on Fast trials and GQI score*
- *Mean latency on Slow trials and GRRS score*
- *Mean latency on Slow trials and GQI score"*

Overall, we find fairly null to weak correlations between individual differences in scores on the Gender Identity Reflection and Rumination Scale (GRRS) and Genderqueer Identity Scale (GQI) and measures of task performance (i.e. accuracy and latency) for all participant groups also see Table 6 below).

Table 6. *Correlations between individual difference and task performance measures*

Participant group	Cisgender Heterosexual		Cisgender LGBPQ+		Gender Diverse	
Measure	GRRS	GQI	GRRS	GQI	GRRS	GQI
Accuracy Correlations						
Fast Trials (2000ms response window)	-.038	-.054	.039	-.007	.065	.180*
Slow Trials (4000ms response window)	.032	.006	-.031	-.003	.113	.089
Latency Correlations						
Fast Trials (2000ms response window)	−.127	−.157*	−.059	−.074	−.089	−.089
Slow Trials (4000ms response window)	−.064	.104	−.015	-.060	-.005	−.047

*p < .05.

We will conduct the following linear regressions on gender diverse and cisgender-heterosexual participant data, predicting mean accuracy or latency on slow or fast trials from GRRS score or GQI score:

- *Mean accuracy on slow trials ~ GRRS * participant group*
- *Mean accuracy on fast trials ~ GRRS * participant group*
- *Mean accuracy on slow trials ~ GQI * participant group*

- *Mean accuracy on fast trials ~ GQI * participant group*
- *Mean latency on slow trials ~ GRRS * participant group*
- *Mean latency on fast trials ~ GRRS * participant group*
- *Mean latency on slow trials ~ GQI * participant group*
- *Mean latency on fast trials ~ GQI * participant group*

Overall, we find that GRRS and GQI do not predict gender diverse or cisgender-heterosexual participants' overall perceptual accuracy regardless of the amount of time given to sort each face. Additionally, there are no consistent differences in mean accuracy or mean latency between cisgender heterosexual and gender diverse participants'. Finally, there is weak evidence to suggest that participants' GRRS & GQI scores may predict latency on fast trials. However, the effects go in opposite directions, suggesting that participants with higher GRRS scores take less time to sort faces, while participants with higher GQI scores take more time to sort faces. For details, see regression Tables 7.1-7.8 below.

Table 7.1 *Predicting Mean Accuracy on Slow Trials from GRRS Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	0.19	[0.17, 0.22]	13.33	< .001
GRRS	0.01	[-0.01, 0.02]	0.65	.519
Participant Group	-0.04	[-0.08, 0.01]	-1.50	.135
GRRS x Participant Group	0.00	[-0.02, 0.02]	0.21	.834

Table 7.2 *Predicting Mean Accuracy on Fast Trials from GRRS Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	0.19	[0.16, 0.22]	13.54	< .001
GRRS	0.01	[-0.01, 0.02]	0.73	.466
Participant group	-0.04	[-0.08, 0.01]	-1.62	.107
GRRS x Participant group	0.00	[-0.02, 0.03]	0.33	.739

Table 7.3 *Predicting Mean Accuracy on Slow Trials from GQI Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	0.20	[0.17, 0.24]	12.49	< .001
GQI	0.00	[-0.01, 0.01]	-0.11	.910

Participant group	-0.05	[-0.12, 0.01]	-1.60	.111
GQI x Participant group	0.01	[-0.01, 0.03]	0.79	.430

Table 7.4 *Predicting Mean Accuracy on Fast Trials from GQI Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	0.21	[0.18, 0.24]	13.18	< .001
GQI	0.00	[-0.02, 0.01]	-0.44	.659
Participant group	-0.07	[-0.14, -0.01]	-2.25	.025*
GQI x Participant group	0.02	[-0.01, 0.04]	1.48	.141

Table 7.5 *Predicting Mean Latency on Slow Trials from GRRS Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	1,404.65	[1,259.68, 1,549.61]	19.05	< .001
GRRS	-67.59	[-150.81, 15.64]	-1.60	.111
Participant group	-73.27	[-308.87, 162.34]	-0.61	.541
GRRS x Participant group	58.00	[-50.65, 166.65]	1.05	.294

Table 7.6 *Predicting Mean Latency on Fast Trials from GRRS Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	1,201.29	[1,119.62, 1,282.97]	28.92	< .001
GRRS	-48.23	[-95.59, -0.87]	-2.00	.046*
Participant group	-74.33	[-209.72, 61.06]	-1.08	.281
GRRS x Participant group	39.85	[-22.65, 102.35]	1.25	.211

Table 7.7 *Predicting Mean Latency on Slow Trials from GQI Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	1,212.15	[1,048.81, 1,375.48]	14.59	< .001

GQI	33.87	[-33.15, 100.89]	0.99	.321
Participant group	241.08	[-86.02, 568.18]	1.45	.148
GQI x Participant group	-76.50	[-182.00, 29.00]	-1.43	.155

Table 7.8 *Predicting Mean Latency on Fast Trials from GQI Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	1,021.13	[930.01, 1,112.25]	22.04	< .001
GQI	42.54	[5.19, 79.88]	2.24	.026*
Participant group	110.96	[-75.31, 297.24]	1.17	.242
GQI x Participant group	-50.29	[-109.98, 9.39]	-1.66	.098

We will repeat the linear regressions listed above on the full dataset, including all three participant groups and using the cis-LGBPQ+ participant group as the reference group.

Rerunning the linear regressions above with cis-LGBPQ+ participants as the reference group yields no significant effects, suggesting that neither individual scores on the GRRS/GQI nor participant identity predict participants' accuracy or latency when sorting faces. For details, see regression tables 8.1-8.8 below.

Table 8.1 *Predicting Mean Accuracy on Slow Trials from GRRS Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	0.19	[0.16, 0.23]	11.97	< .001
GRRS	0.00	[-0.02, 0.01]	-0.49	.623
Cis-Heterosexual (vs. Cis-LGBIQ+)	0.00	[-0.04, 0.04]	0.03	.978
Gender Diverse (vs. Cis-LGBIQ+)	-0.03	[-0.08, 0.01]	-1.42	.155
GRRS Cis-Heterosexual (vs. Cis-LGBIQ+)	0.01	[-0.01, 0.03]	0.81	.418
GRRS Gender Diverse (vs. Cis-LGBIQ+)	0.01	[-0.01, 0.03]	1.09	.276

Table 8.2 *Predicting Mean Accuracy on Fast Trials from GRRS Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	0.18	[0.15, 0.21]	11.32	< .001
GRRS	0.00	[-0.01, 0.02]	0.33	.740
Cis-Heterosexual (vs. Cis-LGBIQ+)	0.01	[-0.03, 0.05]	0.52	.602
Gender Diverse (vs. Cis-LGBIQ+)	-0.03	[-0.07, 0.02]	-1.12	.265
GRRS Cis-Heterosexual (vs. Cis-LGBIQ+)	0.00	[-0.02, 0.03]	0.29	.774
GRRS Gender Diverse (vs. Cis-LGBIQ+)	0.01	[-0.01, 0.03]	0.65	.519

Table 8.3 *Predicting Mean Accuracy on Slow Trials from GQI Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	0.19	[0.15, 0.23]	10.03	< .001
GQI	0.00	[-0.01, 0.01]	-0.17	.867
Cis-Heterosexual (vs. Cis-LGBIQ+)	0.02	[-0.03, 0.06]	0.65	.519
Gender Diverse (vs. Cis-LGBIQ+)	-0.04	[-0.10, 0.03]	-1.08	.281
GQI Cis-Heterosexual (vs. Cis-LGBIQ+)	0.00	[-0.02, 0.02]	0.03	.975
GQI Gender Diverse (vs. Cis-LGBIQ+)	0.01	[-0.01, 0.03]	0.85	.397

Table 8.4 *Predicting Mean Accuracy on Fast Trials from GQI Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	0.19	[0.16, 0.23]	10.59	< .001
GQI	0.00	[-0.01, 0.01]	-0.40	.686
Cis-Heterosexual (vs. Cis-LGBIQ+)	0.02	[-0.03, 0.06]	0.64	.520
Gender Diverse (vs. Cis-LGBIQ+)	-0.06	[-0.12, 0.01]	-1.75	.081

GQI Cis-Heterosexual (vs. Cis-LGBIQ+)	0.00	[-0.02, 0.02]	-0.05	.962
GQI Gender Diverse (vs. Cis-LGBIQ+)	0.01	[0.00, 0.03]	1.50	.133

Table 8.5 *Predicting Mean Latency on Slow Trials from GRRS Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	1,362.55	[1,199.45, 1,525.65]	16.41	< .001
GRRS	-24.59	[-108.20, 59.01]	-0.58	.564
Cis-Heterosexual (vs. Cis-LGBIQ+)	42.10	[-176.22, 260.42]	0.38	.705
Gender Diverse (vs. Cis-LGBIQ+)	-31.17	[-278.50, 216.16]	-0.25	.805
GRRS Cis- Heterosexual (vs. Cis-LGBIQ+)	-42.99	[-161.03, 75.04]	-0.72	.475
GRRS Gender Diverse (vs. Cis-LGBIQ+)	15.01	[-93.99, 124.00]	0.27	.787

Table 8.6 *Predicting Mean Latency on Fast Trials from GRRS Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	1,171.93	[1,077.96, 1,265.90]	24.50	< .001
GRRS	-30.89	[-79.11, 17.33]	-1.26	.209
Cis-Heterosexual (vs. Cis-LGBIQ+)	29.36	[-94.48, 153.20]	0.47	.642
Gender Diverse (vs. Cis-LGBIQ+)	-44.97	[-187.10, 97.17]	-0.62	.535
GRRS Cis- Heterosexual (vs. Cis-LGBIQ+)	-17.34	[-84.51, 49.84]	-0.51	.612
GRRS Gender Diverse (vs. Cis-LGBIQ+)	22.51	[-40.31, 85.34]	0.70	.482

Table 8.7 *Predicting Mean Latency on Slow Trials from GQI Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*

Intercept	1,358.38	[1,168.55, 1,548.22]	14.06	< .001
GQI	-14.32	[-77.12, 48.48]	-0.45	.654
Cis-Heterosexual (vs. Cis-LGBIQ+)	-146.24	[-396.77, 104.30]	-1.15	.252
Gender Diverse (vs. Cis-LGBIQ+)	94.84	[-246.49, 436.18]	0.55	.585
GQI Cis-Heterosexual (vs. Cis-LGBIQ+)	48.19	[-43.70, 140.09]	1.03	.303
GQI Gender Diverse (vs. Cis-LGBIQ+)	-28.31	[-131.24, 74.63]	-0.54	.589

Table 8.8 *Predicting Mean Latency on Fast Trials from GQI Score x Participant Group*

Predictor	*b*	95% CI	*t*	*p*
Intercept	1,048.32	[941.01, 1,155.63]	19.19	< .001
GQI	22.26	[-13.35, 57.87]	1.23	.220
Cis-Heterosexual (vs. Cis-LGBIQ+)	-27.19	[-167.27, 112.88]	-0.38	.703
Gender Diverse (vs. Cis-LGBIQ+)	83.77	[-109.32, 276.86]	0.85	.394
GQI Cis-Heterosexual (vs. Cis-LGBIQ+)	20.28	[-31.00, 71.56]	0.78	.438
GQI Gender Diverse (vs. Cis-LGBIQ+)	-30.01	[-88.19, 28.16]	-1.01	.311

Study 3 exploratory analyses

As in previous studies, we investigated whether or not the accuracy of participants judgements varied by race by running a multilevel regression model on data from [gender] trials predicting error from race of the face, location (distance from the nearest end-point), and their interaction. As before, we also included a random intercept for participant ID, as well as a random slope within-participant for location. Replicating study 1a, we found that there was overall greater error for more intermediary stimuli (b = .106, $t(779) = 15.58$, $p < .001$). Like in Study 1a, we found that that this effect was exacerbated for racial minorities in comparison to White faces, resulting in greater perceptual error for Black ($b = 0.026$, $t(56{,}540{,}030) = 6.098$, $p < .001$) and Asian ($b = 0.012$, $t(56{,}530) = 2.911$, $p = .004$) faces with intermediary gender/sex, relative to that of White faces with intermediary gender/sex. Additionally, we also found that in comparison to White faces, people exhibited overall greater perceptual error categorizing Asian faces ($b = .011$, $t(56{,}540) = 4.574$, $p < .001$), but not Black faces.

Study 3 Brief Summary of Findings

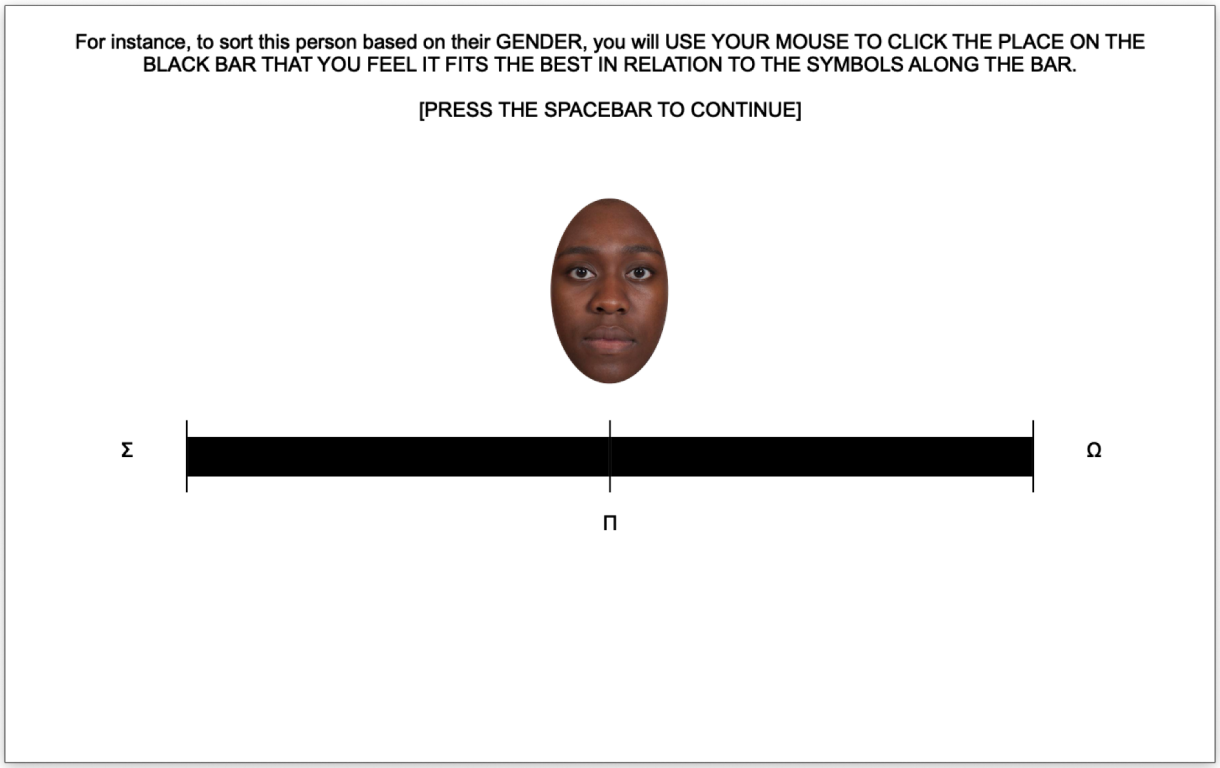
In Study 3, we asked whether people's perceptions of gender/sex differ based on experience. Specifically, we compared accuracy and latency of gender/sex judgments across gender diverse, cisgender LGBPQ+, and cisgender heterosexual participants at both "fast" and "slow" sorting conditions . For accuracy, we found that gender diverse participants across all conditions were more accurate at categorizing faces than cisgender heterosexual and LGBPQ+ participants. Specifically, we found that gender diverse participants were more accurate at categorizing more androgynous faces than both cisgender heterosexual and cisgender LGBPQ+ participants. This supports the possibility that individuals' experience with gender does play a role in gender/sex perception. For latency, we found that participants across all groups were slower at categorizing faces when given a longer window to respond, and that this effect is accentuated when categorizing more androgynous faces. However, there were no differences between groups, which leaves us with inconclusive evidence on the role of response time across groups.

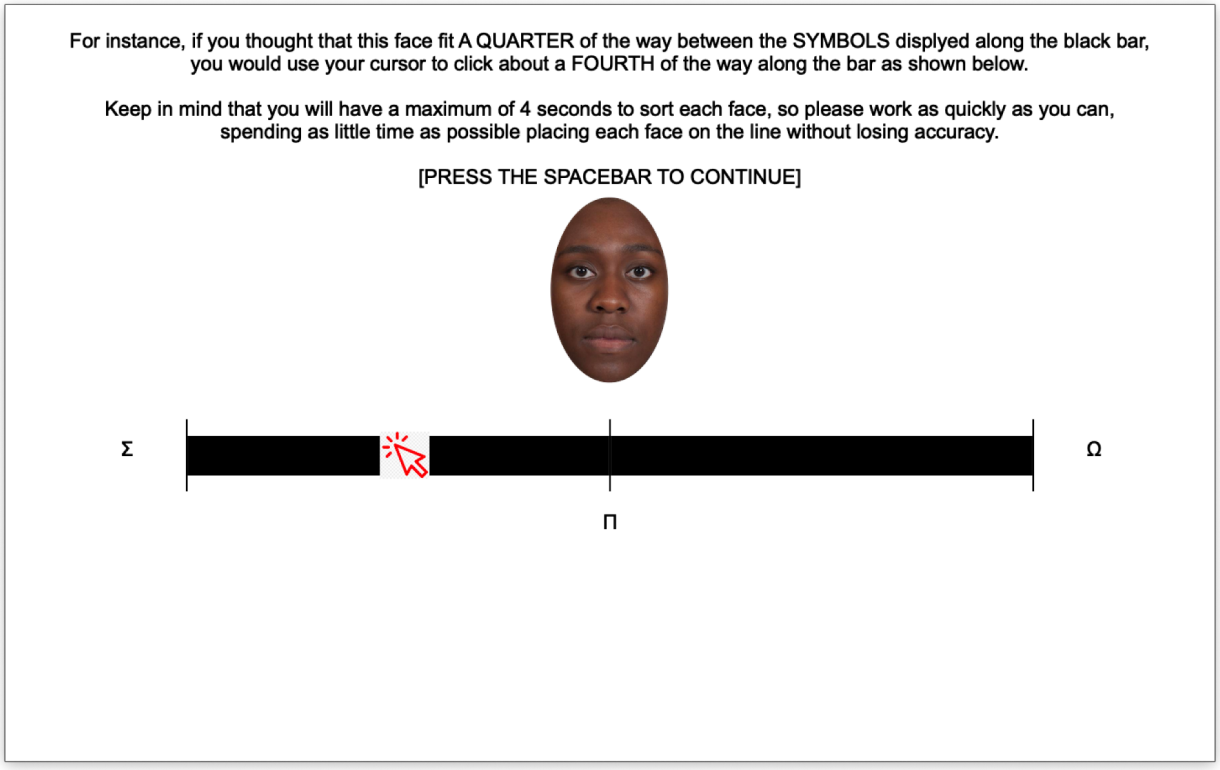
Additional Information about Study 4

Study 4 Detailed Method

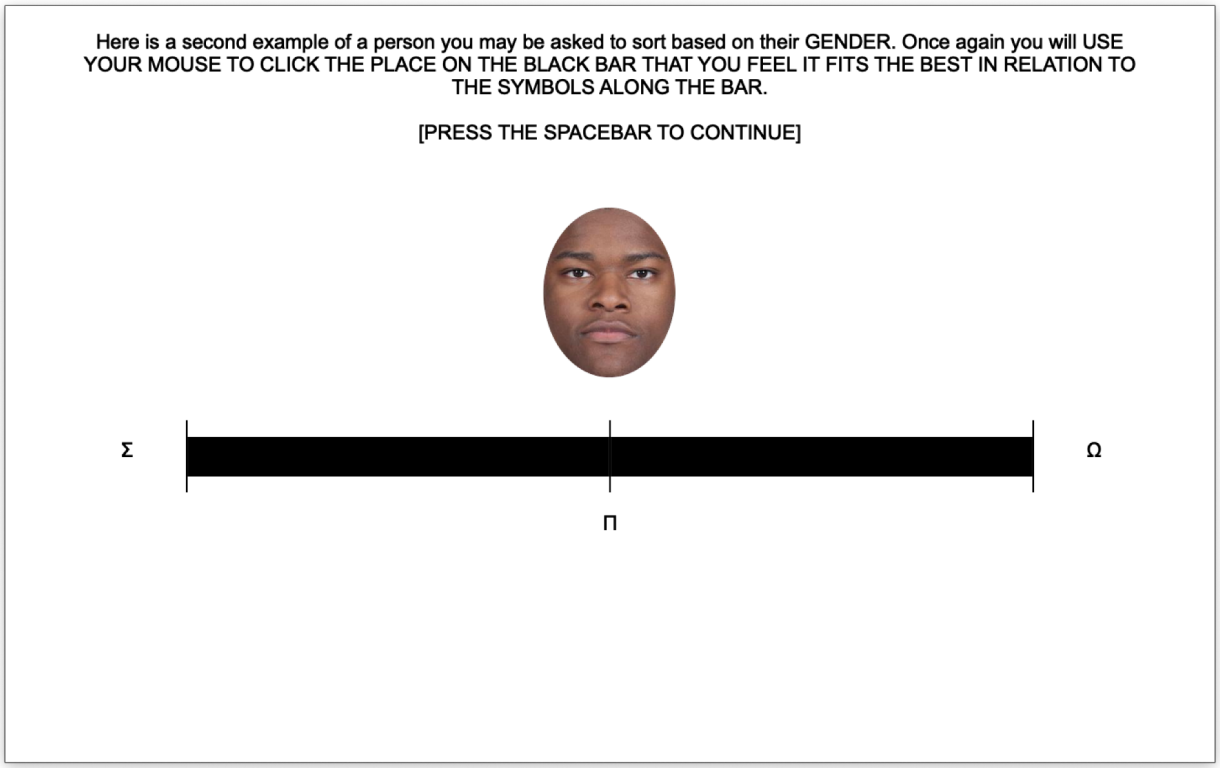
The procedure for Study 4 had two distinct differences from the set-up used in previous studies. First, participants in this study were randomly assigned to complete either a 2-category label or 3-category label version of the task. Second, instead of words (e.g. female or male), all the category labels throughout both conditions were represented by letters in the greek alphabet (e.g. sigma, omega). Our decision to use symbols rather than words was methodological (as opposed to conceptual) because we could not find an adequate word to represent faces between the two endpoints. Before completing the task, participants were told they were being asked to sort people associated with different symbols. Before each block, participants in the 2 category-condition were taught to associate two sets of images with two symbols, each intending to a distinct race or gender category (see Figure 6, Panel A, outer images in manuscript). Participants in the 3 category condition also saw a third set of images and an additional label, intended to represent the place halfway between the two other categories (see Figure 6, Panel A, center image in manuscript). We then presented participants with two example cases, which walked participants through the process of using their cursor to sort faces 25% and 75% of the way along the line between the two symbols (See Figure S1, Panels A and B). Finally, participants were given one last reminder that displayed all labels (as symbols) and associated faces they learned about mapped out on the continuum (See Figure S1, Panel C). At test, participants saw the continuum with all the labels they learned about placed at the expected locations (including in the middle, if in the 3-category condition), and were asked to sort each stimulus along the continuum (see Figure 4 in manuscript). In total, participants completed two blocks, one containing race stimuli and the other gender stimuli in random order each containing 76 trials. Notably, all stimuli used in this study were morphed at a 5% (as opposed to 10%) gradient for finer grained investigation of the intermediate faces (see below for detailed information on stimuli). As in Studies 1b and 2, stimuli were grouped together based on the base pair used to create them, and we counterbalanced the location of labels on either ends so that participants did not associate a label with one particular side of the continuum. Task data were collected on Inquisit Web (Millisecond Software, 2016) and we collected the same variables of interest measured in previous studies. As before, participants were asked to fill out a short survey following completion of the task which included basic demographic information.

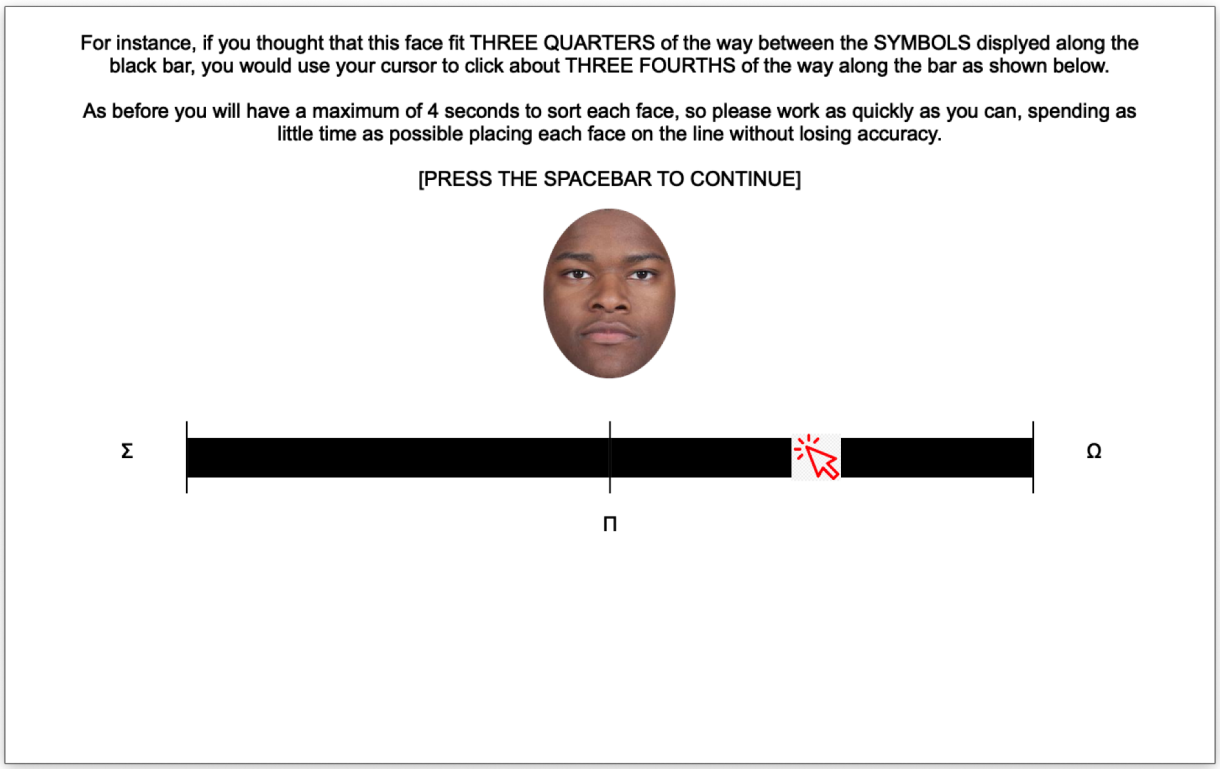
A. Guidance on placing a 25% morph exemplar.





B. Guidance on placing a 75% morph exemplar.





C. Faces and symbols mapped out on continuum

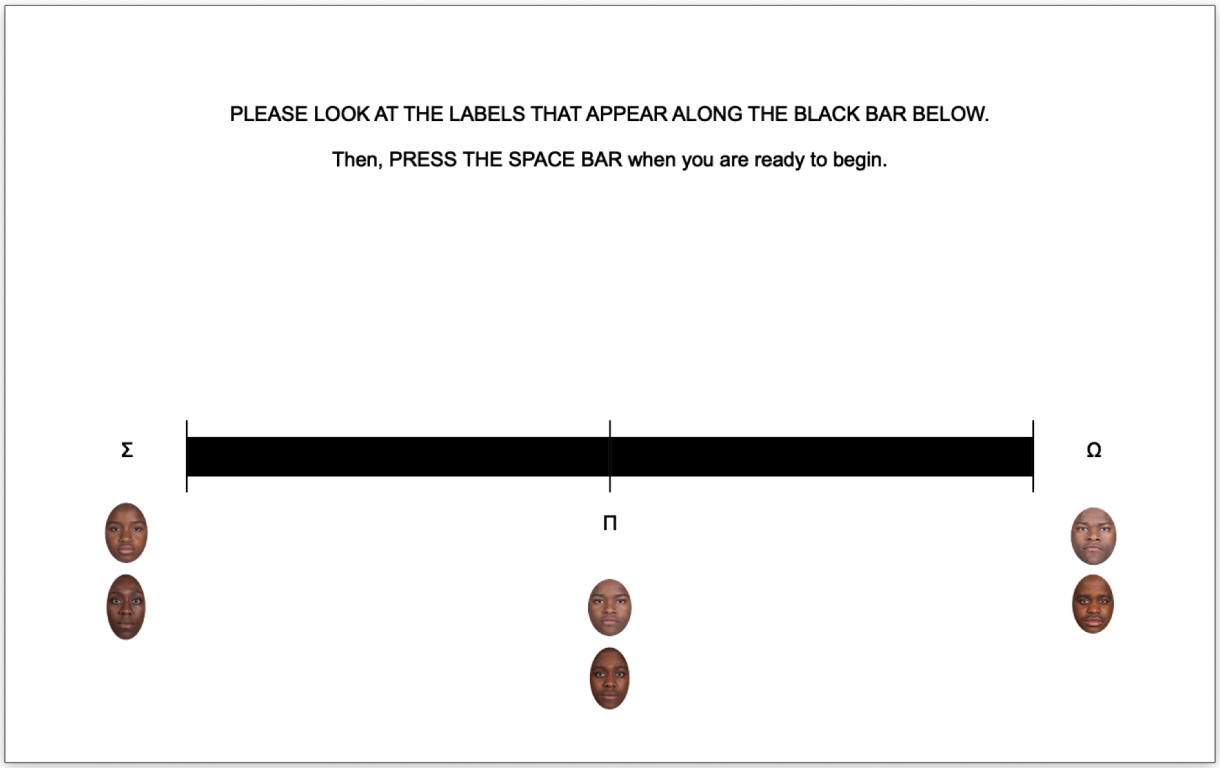


Figure S1. Example training received by participants in the 3-category condition of Study 4

Study 4 Stimuli

Both gender and race stimuli used in Study 4 were created from the same set of base faces used in Study 2. Unlike in previous studies, each of the 12 base pairs were morphed at a 5% gradient in WebMorph to create 19 distinct morphs resulting set of 228 faces for both race and gender, respectively. Participants sorted all morphs from four out of the 12 morphs for both race and gender, resulting in 76 gender stimuli and 76 race stimuli sorted at test.

Study 4 Analyses

Study 4 pre-registered main analyses

"We will run a linear regression predicting accuracy with condition (midpoint vs no midpoint; between-subjects) and ambiguity score (within-subjects) once for the race task and once for the gender task. We will repeat these analyses for latency."

The results from a pair of linear regressions predicting accuracy from condition and ambiguity revealed that participants were overall more accurate at categorizing in the midpoint condition than no-midpoint condition for gender trials, ($b = 3.83$, $t(1) = 6.90$, $p < .001$), but that pattern was reversed for race trials. That is, people were more accurate at categorizing faces in the no-midpoint than the midpoint condition for race ($b = 5.25$, $t(1) = 11.42$, $p < .001$). For both race and gender trials, there was a main

effect of ambiguity such that overall accuracy decreased as ambiguity increased (b = .06, $t(1)$ = 4.18, p < .001 for gender trials and b = .09, $t(1)$ = 8.07, p < .001 for race trials). There was also an interaction between ambiguity on accuracy and condition for both race and gender trials such that the effect of ambiguity was stronger for individuals in the no-midpoint condition than for individuals in the midpoint condition (b = .15, $t(1)$ = 8.24, p < .001 for gender trials and b = .16, $t(1)$ = 10.14, p < .001 for race trials).

A second pair of linear regressions predicting latency from condition and ambiguity revealed that participants in the midpoint condition were slower at categorizing faces in those in the no-midpoint condition (b = 151.98, $t(1)$ = 6.77, p < .001 for gender trials and b = 162.42, $t(1)$ = 7.60, p < .001 for race trials). There was no main effect of ambiguity on latency for either gender (b = .41, $t(1)$ =.76, p = .45) or race (b = .13, $t(1)$ = .25, p = .803). Again, there was an interaction between ambiguity and condition for both race and gender trials such that the effect of ambiguity on latency was stronger for individuals in the no-midpoint condition than for individuals in the midpoint condition (b = 1.64, $t(1)$ = 2.17, p = .030 for gender trials and b = 2.82, $t(1)$ = 3.92, p < .001 for race trials).

"If the interaction is significant, we will run independent samples t-tests for each level of ambiguity comparing the midpoint and no-midpoint trials (using a bonferroni-adjusted p-value of .005). We will repeat these analyses for latency."

Because interactions were significant in all four regression, we ran two-tailed independent-samples t-tests at each level of ambiguity comparing midpoint and no-midpoint trials along the dimensions of accuracy and latency for both gender and race trials (see table below for full reporting). In all subsequent tables, "Ambiguity Level" refers to the difference between a stimuli's morph level and the nearest endpoint. For instance, Ambiguity Level 5 is based on participants' subjective judgments of stimuli that are 5 units away from the endpoints (e.g. the 5% female face/95% male face and the 95% female face/5% male face), Ambiguity Level 10 is based on subjective judgments of stimuli that are 10 units away from the endpoints (e.g. the 10% female face/90% male face and the 90% female face/10% male face), etc. all the way up to Ambiguity Level 50 which is based on subjective judgments of stimuli that are 50 units away from the endpoints (e.g. the 50% female face/50% male face). Note that there are half as many trials at Ambiguity Level 50 because it was the only number without a mirror image across the 50% mark.

Table 9. Study 4 Midpoint vs. No-midpoint planned post-hoc comparisons

Comparison	Midpoint Condition	No-midpoint Condition	99% CI	*t*	*p*
	M	*M*			
Gender Accuracy					
Ambiguity Level 5	17.77	16.15	[-.77, 4.01]	1.75	.080

Ambiguity Level 10	18.44	16.45	[-.36, 4.34]	2.19	.029
Ambiguity Level 15	17.56	16.16	[-.74, 3.55]	1.69	.092
Ambiguity Level 20	18.18	16.53	[-.36, 3.66]	2.12	.035
Ambiguity Level 25	19.25	18.12	[-.76, 3.02]	1.54	.123
Ambiguity Level 30	19.70	18.69	[-.79, 2.80]	1.44	.149
Ambiguity Level 35	20.06	21.24	[-3.01, .66]	-1.65	.100
Ambiguity Level 40	19.53	23.26	[-.74, 3.55]	-5.01	<.001**
Ambiguity Level 45	19.88	23.60	[-5.67, -1.79]	-4.96	<.001**
Ambiguity Level 50	19.85	24.51	[-7.54, -1.78]	-4.18	<.001**
Race Accuracy					
Ambiguity Level 5	14.73	12.49	[.23, 4.26]	2.87	.004**
Ambiguity Level 10	14.98	11.62	[1.55, 5.18]	4.78	<.001**
Ambiguity Level 15	15.89	11.60	[2.73, 5.86]	7.06	<.001**
Ambiguity Level 20	16.99	13.98	[1.39, 4.63]	4.79	<.001**
Ambiguity Level 25	18.00	15.59	[.89, 3.92]	4.10	<.001**
Ambiguity Level 30	18.69	17.24	[-.05, 2.96]	2.50	.013
Ambiguity Level 35	18.67	19.12	[-2.07, 1.18]	-0.70	.482

Ambiguity Level 40	18.67	20.01	[-2.99, .32]	-2.09	.037
Ambiguity Level 45	17.97	20.27	[-4.06, -.54]	-3.37	<.001**
Ambiguity Level 50	17.41	21.53	[-6.75, -5.51]	-4.07	<.001**
Gender Latency					
Ambiguity Level 5	1501.64	1353.53	[61.50, 234.71]	4.81	<.001**
Ambiguity Level 10	1444.20	1316.36	[52.53, 203.15]	4.38	<.001**
Ambiguity Level 15	1422.08	1335.04	[10.99 163.10]	2.95	.003*
Ambiguity Level 20	1456.57	1290.85	[86.16, 245.28]	5.37	<.001**
Ambiguity Level 25	1449.50	1323.95	[43.45, 207.65]	3.94	<.001**
Ambiguity Level 30	1432.19	1333.16	[17.40, 180.66]	3.13	.002*
Ambiguity Level 35	1468.76	1381.52	[.50, 173.99]	2.59	.009*
Ambiguity Level 40	1471.17	1382.20	[.27, 177.65]	2.59	.010
Ambiguity Level 45	1479.94	1413.68	[-24.53, 157.05]	1.88	.060
Ambiguity Level 50	1494.57	1417.82	[-51.93, 205.44]	1.54	.124
Race latency					
Ambiguity Level 5	1473.37	1345.01	[55.58, 201.15]	4.55	<.001**
Ambiguity Level 10	1466.73	1313.85	[79.73, 226.02]	5.39	<.001**

Ambiguity Level 15	1431.17	1328.38	[28.86, 176.72]	3.59	<.001**
Ambiguity Level 20	1450.18	1330.05	[40.79, 199.48]	3.90	<.001**
Ambiguity Level 25	1424.13	1321.75	[27.02, 177.75]	3.50	<.001**
Ambiguity Level 30	1424.53	1351.89	[-7.84, 153.12]	2.33	.020
Ambiguity Level 35	1455.74	1389.59	[-16.51, 148.80]	2.06	.039
Ambiguity Level 40	1473.08	1400.91	[-9.79, 154.13]	2.27	.023
Ambiguity Level 45	1464.38	1449.36	[-69.78, 99.81]	0.46	.648
Ambiguity Level 50	1467.85	1459.28	[-110.30, 127.43]	0.19	.852

Table 9 contains full reporting for planned post-hoc t-tests for Study 4. Accuracy comparisons compare to the average amount of error between Midpoint vs. No-Midpoint conditions at a given ambiguity level, while latency comparisons compare the average amount of time it took participants to sort stimuli between the conditions at a given ambiguity level. Ambiguity level is the actual distance of each stimulus from the the nearest endpoint of the scale. For instance, Ambiguity Level 5 can refer either to a 5% female face/95% male face or a 95% female face/5% male face for gender comparisons. ** $p < .001$, * $p < .005$.

Study 4 pre-registered additional analyses

"We will calculate and report the number of individuals showing a negative correlation between objective [location] and subjective judgment while completing each task. For each participant, we will calculate 2 correlations (for gender and race phenotype trials, separately), and then exclude all data from tasks where participants' subjective judgments do not positively correlate with [objective location] before running our main analyses."

Out of 226 participants with eligible race task data after initial exclusions based on careless responding, 8 people showed a negative correlation between objective location and subjective judgment while 218 showed a positive correlation. When including all 226 participants the mean pearson correlation between objective location and subjective judgment was $r = .730$, but after excluding the 8 participants with a negative correlation, the mean pearson correlation of the remaining 218 participants was $r = .771$.

Out of 219 participants with eligible gender task data after initial exclusions based on careless responding, 16 people showed a negative correlation between objective location and subjective judgment

while 203 showed a positive correlation. When including all 219 participants the mean pearson correlation between objective location and subjective judgment was $r = .596$, but after excluding the 16 participants with a negative correlation, the mean pearson correlation of the remaining 203 participants was $r = .673$. As negative correlations likely indicated that people had reversed the category-symbol pairings, these participants were dropped, in line with the preregistration (previous studies did not involve learning categories and did not have participants with negative correlations).

Study 4 exploratory analyses

Like all previous studies, we investigated whether or not the accuracy of participants judgements varied by race by running a multilevel regression model on data from [gender] trials predicting error from race of the face, location (distance from the nearest end-point), and their interaction. As before, we also included a random intercept for participant ID, as well as a random slope within-participant for location. However, unlike all previous studies, we found no evidence of differences in participants' judgements by race.

Study 4 Brief Summary of Findings

In Study 4, we asked how an additional category label at the midpoint of the continuum influenced latency and accuracy in subjective judgments of race and gender. For latency, an additional label at the midpoint made participants slower overall. This finding is in line with previous literature on cognitive load, which has found that additional information slows down cognitive processing (Sweller, 1988). For accuracy, effects were mixed. An additional label was associated with increased accuracy when perceiving gender but not race. However, for both race and gender perception, providing an additional label at the midpoint increased participants' accuracy at categorizing faces with objective locations at the midpoint. This suggests that the addition of a third category improved accuracy for the most intermediary morphs.

Additional Information about Study 5

Study 5 Detailed Method

The procedure for Study 5 was similar to that of Study 4 with one key difference. In addition to the 2-category label and 3-category label version of the task, we also introduced a third condition that was identical to the 2-category condition except that participants in this condition sorted stimuli on a black line with a demarcation (but no label) at the midpoint (henceforth, demarcation condition). As before, participants were instructed to associate images with symbols as appropriate based on condition (with participants in the demarcation condition seeing only two labels, just like participants in the 2-category condition). At test, participants saw a black line with 2 labels (2-category condition), 3 labels (3-category condition), or 2 labels plus a demarcation in the middle (demarcation condition), and were asked to sort each stimulus along the line in relation to the labels at either end (see Figure 6; Panel B in manuscript). Unlike Study 4, participants completed just a single block of 76 trials sorting White faces along the dimension of gender. As in previous studies, stimuli were grouped together based on the base pair used to create them, the location of labels was counterbalanced, task data were collected on Inquisit Web (Millisecond Software, 2016), and we collected the same variables of interest measured in previous

studies. Finally, participants were asked to fill out a short demographic questionnaire following completion of the task.

Study 5 Stimuli

Stimuli for Study 5 consisted of the White faces used as gender stimuli in Study 4. These 76 faces were created from 4 base pairs morphed at a 5% gradient using the "transform" tool in the WebMorph interface. At test, participants sorted all 76 morphs in a single testing block.

Study 5 Analyses

Study 5 pre-registered main analyses

"We will run a linear regression predicting accuracy with condition (2-category vs. 3-category vs. demarcation; between-subjects) and ambiguity score (within-subjects)."

A linear regressions predicting accuracy from condition and ambiguity replicated Study 4 such that that participants were overall more accurate in the 2-category condition than 3-category condition, (b = 8.06, $t(1) = 14.10$, $p < .001$) and the demarcation condition ($b = 3.77$, $t(1) = 6.64$, $p < .001$). Similarly, participants were overall more accurate in the demarcation condition than the 3-category condition (b = 4.289, $t(1) = 7.42$, $p < .001$). Like in Study 4, there was a main effect of ambiguity such that participants were less accurate at categorizing more ambiguous faces ($b = .20$, $t(1) = 14.88$, $p < .001$). Finally, there was also an interaction between ambiguity and condition such that the negative impact of ambiguity on accuracy was stronger for participants in the 2-category condition than either the demarcation condition (b = -.13, $t(1) = -6.91$, $p < .001$) or 3-category condition ($b = -.28$, $t(1) = -14.94$, $p < .001$).

"To further understand these interactions (if they are found), we will run independent samples t-tests for each level of ambiguity comparing the 2-category vs. 3-category conditions, the demarcation vs. 2-category condition, and the demarcation vs. 3 category condition (using a bonferroni-adjusted p-value of .0008)."

Because interactions were significant, we ran two-tailed independent-samples t-tests at each level of ambiguity comparing the 2-category vs. 3-category conditions, demarcation vs. 2-category conditions, and demarcation vs. 3-category conditions along the dimensions of accuracy and latency (see table below for full reporting).

Study 5 pre-registered additional analyses

"We will calculate and report the number of individuals showing a negative correlation between objective [location] and subjective judgment while completing each task. (We will calculate a correlation for each participant, and then exclude all data from tasks where participants' subjective judgments do not positively correlate with objective locations before running our main analyses.)"

Out of 112 participants in the 3-category condition with eligible task data after initial exclusions based on careless responding, 8 people showed a negative correlation between objective location and subjective judgment while 104 showed a positive correlation. When including all 112 participants the mean pearson correlation between objective location and subjective judgment was $r = .560$, but after

excluding the 8 participants with a negative correlation, the mean pearson correlation of the remaining 104 participants was $r = .609$.

Out of 117 participants in the 2-category condition with eligible task data after initial exclusions based on careless responding, 5 people showed a negative correlation between objective location and subjective judgment while 112 showed a positive correlation. When including all 117 participants the mean pearson correlation between objective location and subjective judgment was $r = .646$, but after excluding the 5 participants with a negative correlation, the mean pearson correlation of the remaining 112 participants was $r = .691$.

Out of 116 participants in the demarcation condition with eligible task data after initial exclusions based on careless responding, 9 people showed a negative correlation between objective location and subjective judgment while 107 showed a positive correlation. When including all 116 participants the mean pearson correlation between objective location and subjective judgment was $r = .585$, but after excluding the 9 participants with a negative correlation, the mean pearson correlation of the remaining 107 participants was $r = .650$.

"We will also calculate latency (i.e. the length of time in milliseconds it takes for participants to sort each item) for each trial, and run a linear regression predicting latency with condition (2-category vs. 3-category vs. demarcation; between-subjects) and ambiguity score (within-subjects). Again, we will follow up with pairwise comparisons at each level of ambiguity (using a bonferroni-adjusted p-value of .0008)."

A linear regression predicting latency from condition and ambiguity revealed that participants in the 3-category condition were slower at categorizing faces in those in the 2-category condition ($b = 70.02$, $t(1) = 3.30$, $p < .001$), but there no difference for between the demarcation and 2-category conditions ($b = -17.19$, $t(1) = -0.82$, $p < .414$). There was a main effect of ambiguity on latency such that people took longer to categorize more ambiguous faces ($b = 3.82$, $t(1) = 7.71$, $p < .001$). Additionally, there was an interaction between ambiguity and condition such that the effect of ambiguity on latency was stronger for individuals in the 2-category condition than for individuals in the 3-category condition ($b = -2.04$, $t(1) = -2.85$, $p = .004$), but not those in the demarcation condition ($b = .73$, $t(1) = 1.04$, $p = .301$).

Table 10. Study 5 3-category vs. 2-category planned post-hoc comparisons

Comparison	3-Category Condition	2-Category Condition	99% CI	*t*	*p*
	M	*M*			
Accuracy					
Ambiguity Level 5	21.42	15.72	[2.90, 8.49]	5.25	<.0008*
Ambiguity Level 10	20.83	15.77	[2.47, 7.66]	5.04	<.0008*

Ambiguity Level 15	20.12	15.66	[2.15, 6.77]	4.99	<.0008*
Ambiguity Level 20	19.72	16.56	[1.05, 5.28]	3.86	<.0008*
Ambiguity Level 25	18.30	17.53	[-1.11, 2.66]	1.06	.2898
Ambiguity Level 30	19.27	19.29	[-1.85, 1.81]	-.03	.9797
Ambiguity Level 35	18.27	20.58	[-4.13, -.50]	-3.29	.0010
Ambiguity Level 40	18.05	22.35	[-6.05, -2.54]	-6.31	<.0008*
Ambiguity Level 45	17.85	22.82	[-6.82, -3.12]	-6.92	<.0008*
Ambiguity Level 50	17.14	22.97	[-8.49, -3.17]	-5.65	<.0008*
Latency					
Ambiguity Level 5	1412.86	1375.54	[-30.67, 105.32]	1.42	.1571
Ambiguity Level 10	1425.86	1356.39	[-3.51, 142.45]	2.45	.0142
Ambiguity Level 15	1377.55	1355.42	[-48.81, 93.07]	.80	.4213
Ambiguity Level 20	1415.69	1371.23	[-29.34, 118.28]	1.55	.1205
Ambiguity Level 25	1385.74	1362.57	[-53.56, 99.92]	.78	.4362
Ambiguity Level 30	1428.40	1389.24	[-41.88, 120.20]	1.25	.2129
Ambiguity Level 35	1423.28	1446.77	[-102.60, 55.63]	-.77	.4441

Ambiguity Level 40	1449.67	1461.02	[-95.33, 72.64]	-.35	.7276
Ambiguity Level 45	1499.67	1524.69	[-110.07, 60.04]	-.76	.4483
Ambiguity Level 50	1474.72	1515.03	[-161.91, 81.30]	-.86	.3925

Table 10 contains full reporting for planned post-hoc t-tests for Study 5 comparing the mean accuracy and latency of participants' subjective judgments at each level of ambiguity in the 3-Category vs. 2-Category condition.

Table 11. Study 5 Demarcation vs. 2-category planned post-hoc comparisons

Comparison	Demarcation Condition	2-Category Condition	99% CI	*t*	*p*
	M	*M*			
Accuracy					
Ambiguity Level 5	19.39	15.72	[1.02, 6.33]	3.57	<.0008*
Ambiguity Level 10	17.96	15.77	[-.30, 4.69]	2.26	.0237
Ambiguity Level 15	17.48	15.66	[-.38, 4.02]	2.13	.0332
Ambiguity Level 20	17.63	16.56	[-1.03, 3.18]	1.32	.1878
Ambiguity Level 25	17.91	17.53	[-1.53, 2.30]	.52	.6028
Ambiguity Level 30	18.58	19.29	[-2.53, 1.11]	-1.00	.3165
Ambiguity Level 35	20.05	20.58	[-2.38, 1.31]	-.75	.4521
Ambiguity Level 40	19.85	22.35	[-4.30, -.71]	-3.59	<.0008*
Ambiguity Level 45	21.08	22.82	[-3.58, .10]	-2.44	.0149

Ambiguity Level 50	21.20	22.97	[-4.38, .85]	-1.74	.0817
Latency					
Ambiguity Level 5	1362.36	1375.54	[-81.37, 55.01]	-.50	.6182
Ambiguity Level 10	1366.34	1356.39	[-59.85, 79.75]	.37	.7133
Ambiguity Level 15	1328.23	1355.42	[-97.40, 43.01]	-1.00	.3179
Ambiguity Level 20	1358.78	1371.23	[-87.32, 62.42]	-.43	.6681
Ambiguity Level 25	1384.24	1362.57	[-55.40, 98.75]	.73	.4684
Ambiguity Level 30	1405.62	1389.24	[-62.38, 95.14]	.54	.5918
Ambiguity Level 35	1405.91	1446.77	[-121.07, 39.36]	-1.31	.1892
Ambiguity Level 40	1492.22	1461.02	[-52.54, 114.95]	.96	.3367
Ambiguity Level 45	1544.93	1524.69	[-68.30, 108.79]	.59	.5555
Ambiguity Level 50	1543.86	1515.03	[-95.55, 153.21]	.60	.5497

Table 11 contains full reporting for planned post-hoc t-tests for Study 5 comparing the mean accuracy and latency of participants' subjective judgments at each level of ambiguity in the Demarcation vs. 2-Category condition. . $^{**}p < .0001$, $^{*}p < .0008$.

Table 12. Study 5 Demarcation vs. 3-category planned post-hoc comparisons

Comparison	Demarcation Condition	3-Category Condition	99% CI	*t*	*p*
	M	*M*			
Accuracy					

Ambiguity Level 5	19.39	21.42	[-4.82,.78]	-1.86	.0627
Ambiguity Level 10	17.96	20.83	[-5.39, -.36]	-2.95	.0033
Ambiguity Level 15	17.48	20.12	[-4.91, -.36]	-2.99	.0028
Ambiguity Level 20	17.63	19.72	[-4.19, .003]	-2.58	.0101
Ambiguity Level 25	17.91	18.30	[-2.28, 1.51]	-.53	.5974
Ambiguity Level 30	18.58	19.27	[-2.49, 1.12]	-.99	.3226
Ambiguity Level 35	20.05	18.27	[-.02, 3.57]	2.56	.0107
Ambiguity Level 40	19.85	18.05	[.02, 3.56]	2.62	.0090
Ambiguity Level 45	21.08	17.85	[1.37, 5.09]	4.48	<.0001**
Ambiguity Level 50	21.20	17.14	[1.42, 6.71]	3.97	<.0001**
Latency					
Ambiguity Level 5	1362.36	1412.86	[-121.97, 20.97]	-1.82	.0686
Ambiguity Level 10	1366.34	1425.86	[-133.88, 14.83]	-2.06	.0391
Ambiguity Level 15	1328.23	1377.55	[-121.03, 22.37]	-1.77	.0762
Ambiguity Level 20	1358.78	1415.69	[-133.18, 19.34]	-1.92	.0544
Ambiguity Level 25	1384.24	1385.74	[-80.16, 77.16]	-.05	.9608
Ambiguity Level 30	1405.62	1428.40	[-105.03, 59.47]	-.71	.4752

Ambiguity Level 35	1405.91	1423.28	[-99.08, 64.33]	-.55	.5835
Ambiguity Level 40	1492.22	1449.67	[-40.88, 125.99]	1.32	.1886
Ambiguity Level 45	1544.93	1499.67	[-42.36, 132.88]	1.33	.1830
Ambiguity Level 50	1543.86	1474.72	[-51.36, 189.63]	1.48	.1389

Table 12 contains full reporting for planned post-hoc t-tests for Study 5 comparing the mean accuracy and latency of participants' subjective judgments at each level of ambiguity in the Demarcation vs. 2-Category condition. ** $p < .0001$, * $p < .0008$

Study 5 Brief Summary of Findings

Study 5 served to replicate findings in study three, and further elucidate the role of the midpoint label in people's perceptual processing. As before, we found that when making categorizations about face gender, participants in the 3-category condition were slower but more accurate at categorizing faces than those in the 2-category condition. We also replicated our finding that participants in the 3-category condition were more accurate at categorizing the most intermediary faces than those in the 2-category condition. Because the Demarcation condition was new to his study, we did additional analyses to measure how it would compare to the 2 and 3 category conditions. First, we found no overall differences in latency between the demarcation condition and either the 2-category or 3-category conditions. For accuracy, we found that participants in the demarcation condition were *more* accurate than those in the 2-category condition, but *less* accurate than those in the 3-category condition. Additionally, we found that participants in the 3-category condition were significantly more accurate at sorting intermediary morphs than either those in the demarcation and 2-category group (and that the latter two groups themselves did not significantly different from one another). Taken together, this final study affirms the idea that having a labeled midpoint category (like participants in the 3 category condition) provided participants with useful perceptual information above and beyond visual saliency.