AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 12, 2931–2950
https://doi.org/10.1037/xge0001512

Flexibility in Continuous Judgments of Gender/Sex and Race

S. Atwood[1, 2], Dominic J. Gibson[2], Sofía Briones Ramírez[1], and Kristina R. Olson[1, 2]
[1] Department of Psychology, Princeton University
[2] Department of Psychology, University of Washington

Across six preregistered studies ($N = 1{,}292$; recruited from university subject pools and Prolific Academic), we investigate how face perception along the dimensions of gender/sex and race can vary based on immediate contextual information as well as personal experience. In Studies 1a and 1b, we find that when placing stimuli along a continuum from male to female, cisgender participants sort prototypical gender/sex faces in a bimodal fashion and show less consensus and greater error when placing faces of intermediate gender/sex. We replicate and extend these findings to race in Study 2. In Study 3, we test whether sorting patterns can be influenced by preexisting experiences, and find evidence that transgender/nonbinary participants show less error than cisgender heterosexual participants when sorting intermediary faces. Finally, in Studies 4 and 5, we test whether cisgender participants' judgments of intermediary faces along the continuum are influenced by the specific circumstances under which they are asked to sort. Here, we find that changing the sorting framework to include a third category resulted in less error when placing intermediary faces along the continuum than when participants were provided with only two category labels or two categories and a line at the midpoint, suggesting that new perceptual categories introduced with minimal training can be adopted quickly and successfully in a perceptual task. These data suggest that both long-term life experiences and quick experimental interventions can shape how we think about gender/sex and race.

Public Significance Statement
As scientists and as a society, we often describe social dimensions categorically, using terms like male and female, man and woman, or Black, Asian, and White. Yet in reality, these are continuous dimensions wherein people can be intersex, nonbinary, multiracial, or use other terms to describe themselves and others. This work investigates gender/sex and race face perception, with a focus on how less discrete, more intermediate faces are perceived. Across six studies, we find that people are influenced by both well-known labels (e.g., Black, White, man, and woman) as well as personal experience when sorting faces that vary by gender/sex and race. This both reaffirms the importance of category labels in our thinking about social categories and suggests that these categories might be more flexible than previously acknowledged.

Keywords: social cognition, categorization, face perception, gender/sex, race

Supplemental materials: https://doi.org/10.1037/xge0001512.supp

This article was published Online First February 8, 2024.
S. Atwood https://orcid.org/0000-0003-2889-9575
Dominic J. Gibson https://orcid.org/0000-0001-8841-2840
Kristina R. Olson https://orcid.org/0000-0002-1111-544X
Subsets of these data have been presented at the Annual Meeting of the Society for Personality and Social Psychology, the Trends in Psychology Summit, the Person Memory Interest Group, the 6 College Conference, and at social psychology seminars at the University of Washington and at Princeton University. This work has not been previously published, nor is it under consideration for publication elsewhere. The authors wish to thank Natalie Gallagher and Robin Sifre for consulting on analyses, as well as Eric Hehman for generous comments on previous drafts. This research was supported by National Science Foundation Grant SMA-1837857/SMA-2041463 awarded to Kristina Olson and National Science Foundation Graduate Research Fellowship awarded to S. Atwood. The authors declare no financial conflicts or competing interests that arise from this research or its publication. All data, study materials, code, and supplementary material are available at the project page on the Open Science Framework (https://osf.io/3xzq9/).

S. Atwood and Dominic J. Gibson have contributed equally to this article. S. Atwood served as lead for data curation, formal analysis, project administration, software, and visualization, contributed equally to investigation, and served in a supporting role for funding acquisition. Dominic J. Gibson contributed equally to formal analysis and served in a supporting role for supervision. Sofía Briones Ramírez served in a supporting role for conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, and writing–review and editing. Kristina R. Olson served as lead for funding acquisition and supervision. S. Atwood, Dominic J. Gibson, and Kristina R. Olson contributed equally to conceptualization, writing–review and editing, and methodology. S. Atwood and Dominic J. Gibson contributed equally to writing–original draft.

Correspondence concerning this article should be addressed to S. Atwood, Department of Psychology, Princeton University, Peretsman Scully Hall, Princeton, NJ 08540, United States. Email: satwood@princeton.edu

How do people think about and use social dimensions like gender/sex[1] and race in our social judgments? With increased numbers of people identifying as nonbinary (Wilson & Meyer, 2021), multiracial (Shih & Sanchez, 2009), and intersex (Davis, 2015; Reis, 2021), there is growing recognition that human diversity is not captured by mutually exclusive and distinct categories (e.g., Crisp & Hewstone, 2007; Dunham & Olson, 2016; Gaither, 2015; Hyde et al., 2019; Kang & Bodenhausen, 2015). While we can intellectually recognize the continuous nature of these dimensions of human variation, it is less clear whether people treat variation in gender/sex and race, at least as connoted by faces, as continuous or whether people instead think of this variation as fitting within discrete social categories.

Previous research suggests that when making judgments about the gender/sex or race of faces, people rely heavily on discrete social categories (for race see Levin & Angelone, 2002; MacLin & Malpass, 2001; Walker & Tanaka, 2003; for gender/sex see Campanella et al., 2001; Freeman, Rule, et al., 2010). For instance, past work has found that when categorizing morphed faces that transitioned gradually from a male to female face, participants' categorizations abruptly shifted from overwhelmingly male to overwhelmingly female (Campanella et al., 2001). However, most of this work asks participants to sort stimuli into dichotomous or discrete categories rather than allowing participants to make judgments along a continuum. Thus, much of the past work provides an incomplete picture of how people think about variation in race or gender/sex within a social category and does not reveal the extent to which difficulty sorting intermediary faces (faces that fall between two categories) is driven by participants' own tendencies to rely on binary categories or the forced-choice nature of the tasks.

To the extent that people's ideas about social dimensions are biased toward discrete categories, it is not clear whether this is especially true of social dimensions or comparable to the role that categories play in ideas about other continuous dimensions. For example, people tend to treat color as categorical (see Harnad, 1987) but people regularly make nuanced judgments about color. Even when placing numbers on a number line, people often anchor their estimates to certain reference points such as the endpoints or midpoint, and yet people routinely make nuanced judgments about quantities (Zax et al., 2019). Do social categories exhibit similarly subtle influences on how people think about individuals along social dimensions or are social categories playing an even greater role in social domains (possibly because they are more complex, socially constructed categories)?

Perhaps most importantly, previous research does not reveal how social categories get their privileged status in shaping how we think about social dimensions like race and gender/sex. Do they reflect a clustering of the phenomena themselves in our social environments (i.e., faces are sexually dimorphic)? Are social categories more malleable and able to be shifted with specific experiences?

In the current work, we sought insight into these questions by asking participants to make judgments[2] about gender/sex (our primary focus) and race along a continuous spectrum and comparing the patterns of responses to how participants make judgments about nonsocial dimensions (i.e., color and number; Studies 1 and 2). Next, we explored possible mechanisms through which participants' social judgments become aligned with distinct social categories by investigating how participants' sorting patterns are related to differences in their own identity/extended experience (Study 3) and/or their immediate experience in the context of the experiment (Studies 4 and 5).

Attention to Gender/Sex and Race

Why might discrete categories play an outsized role in how people think about social dimensions like gender/sex and race? One reason is that social categories are consequential for how we think about and interact with one another across the lifespan. Social categorization colors everything from our childhood playmate preferences (e.g., Fishbein & Imai, 1993; Ruble et al., 2007), to hiring decisions (e.g., Bertrand & Mullainathan, 2004; Olian et al., 1988; Pager & Shepherd, 2008), social relationships (e.g., Eagly & Wood, 1999; Williams & Eberhardt, 2008), and how we treat and are treated by others throughout our lifetime (e.g., Allport, 1954; Ridgeway, 2011).

Accordingly, gender/sex and race are highly salient and appear to be encoded automatically by children and adults (Devine, 1989; Fiske & Neuberg, 1990; Weisman et al., 2015). By adulthood, social categorization of faces occurs as early as 100 ms (in the case of race) or 150 ms (in the case of gender/sex) after seeing a face (Ito & Urland, 2003). Gender/sex categories exist in all known human societies around the world (Bussey, 2011; Eagly & Wood, 1999). Infants differentiate members of gender/sex groups as early as 3 months old (Quinn et al., 2002). Moreover, between 3 and 5 years, children come to understand gender/sex as an enduring and essential part of identity (Ruble et al., 2007; Taylor et al., 2009).

Despite the salience of social categories, determining where individuals fall along social dimensions relative to one another (especially within categories) may be inherently more difficult than making analogous judgments about nonsocial dimensions like color or number. Colors within the visible spectrum can be defined simply by their wavelength and numbers increase, by definition, linearly. In contrast, where a given face falls along social dimensions compared to another face could be determined by many features (e.g., expression, skin tone, size, specific facial features). This multidimensionality may mean that social dimensions are more intrinsically and even arbitrarily defined by discrete categories.

Categorical Perception of Gender/Sex and Race[3]

Due to these or other factors, people appear to rely on discrete categories even when sorting faces that vary continuously (Levin,

[1] To acknowledge the difficulties of firmly distinguishing the concepts of sex and gender and identifying whether one or both are implicated in this area of research, we use the term gender/sex (van Anders, 2015).

[2] In all studies, we calculated participants' per-trial judgment "error" in order to facilitate comparison of participants' categorizations across domains, conditions, and types of tasks. Error referred to the distance from the mathematical location according to stimuli creation. This concept of "error" (as well as accuracy) as we use it, should not be generalized outside the laboratory because there is no "objective" indicator of where one falls on a continuum of gender/sex or race given that these are constructs defined by societal norms that change over time (for additional information on how error is calculated, see the task scoring under the Method section in Study 1a).

[3] There are ongoing debates about whether social perception research of the type discussed in this article is or is not actually measuring "perception" as that term is used outside of social psychology (Firestone & Scholl, 2015). We have largely avoided using this term to describe our work to make clear that we are not assessing low-level visual perception nor do we think the processes involved are those traditionally considered "perception." However, we occasionally use a qualified version of this word to refer to subfields or field-specific terms including "race perception," "person perception," and "face perception" to make clear to our readers, the connection between our work and the work in those subfields.

1996; Levin & Angelone, 2002; Levin & Beale, 2000; for reviews, see Pauker et al., 2010; Pauker, Meyers, et al., 2018; Young et al., 2021). Campanella and colleagues (2001) presented participants with five levels of morphs between male and female faces. When asked to classify each face as male or female, participants nearly always agreed about labeling the more prototypical faces, but were at chance and especially slow at sorting the most intermediary faces. Since participants were only given two options in this study, to classify faces as male or female, their chance performance can be interpreted in multiple ways. Were participants unsure how to categorize the most intermediary faces, perhaps seeing them as androgynous and therefore randomly selecting the male versus female options? Or did half of the participants (or all of the participants on half of the trials) see the faces as male and the other half female?

Similarly, people are both slower and less likely to categorize faces as multiracial (vs. monoracial, see Chen & Hamilton, 2012; Chen et al., 2014) and have poorer memory for faces with ambiguous races (Pauker et al., 2009), suggesting uncertainty and a lack of attention around the use of the multiracial category and its members (Blascovich et al., 1997; Freeman, Pauker, et al., 2010). Still, these patterns could be a function of the tasks participants were asked to complete—which required participants to make categorical judgments and provided no option for more continuous responding. Yet even when they have the opportunity for less discrete options, people sometimes provide discrete labels. For example, when given the opportunity to provide open-ended responses, Nicolas and colleagues (2019) found that multiracial faces (e.g., Black and White biracials) were often given discrete categorical ethnic or religious labels (e.g., Hispanic, Muslim).

Freeman and colleagues allowed participants to give more nuanced responses on a face perception task, by asking them to rate faces morphed by gender/sex along a 9-point scale from *extremely male* to *extremely female* (Freeman, Rule, et al., 2010). In line with Campanella and colleagues (2001), participants' responses were biased toward the endpoints of the scale. While the most androgynous faces were, on average, accurately placed toward the center of the scale, it is still unclear how to interpret the average. For example, did half of the participants see these faces as male and half as female, resulting in an androgynous midpoint average, or were people in agreement about their androgyny? This work leaves open the question of how more androgynous faces are perceived.

Possible Mechanisms

Several possible mechanisms could explain how people's understanding of individuals along social dimensions comes to be aligned with dominant social categories. One possibility is that people may actually encounter more faces that fall toward the ends of social dimensions like race or gender/sex than faces that fall in the middle and thus become more attuned to the types of faces that they see more commonly. For example, humans, like many animals, exhibit sexual dimorphism and this is evident in human facial features (Samal et al., 2007). This account is supported by a body of research that suggests social face perception is shaped by people's social environments (Brigham & Malpass, 1985) and experiences (de Heering & Rossion, 2008).

While some types of experience with faces may be more or less universal, there is also significant variation in people's experience with faces regarding gender/sex and race, and this variation could also impact how people think about or use social categories in their judgments. One prominent example of the influence of personal experience and environment in the domain of race is that people are better at recognizing faces of their own race than faces of other races (Brigham & Barkowitz, 1978; Malpass & Kravitz, 1969; Meissner & Brigham, 2001; Ma et al., 2011). Various, not mutually exclusive, explanations have been offered for the own-race advantage, including the theory that greater exposure to people of one's own race may lead to greater perceptual expertise for processing and remembering own-race faces (Brigham & Malpass, 1985; Carroo, 1986; Correll et al., 2021; Fallshore & Schooler, 1995; Golby et al., 2001; Lindsay et al., 1991; Tanaka et al., 2004). Another possibility is that people are more motivated to individuate own-race faces compared to other race faces, which again leads to better recognition and memory of own-race faces (Hugenberg et al., 2010; Levin & Beale, 2000; Meissner et al., 2005; Simon et al., 2023). While a number of factors might influence when the own-race advantage is likely to be observed, the basic pattern has been demonstrated across multiple methodologies and populations including looking time and/or eye tracking in young infants (Gaither et al., 2012; Sugden & Marquis, 2017), using functional magnetic resonance imaging methodology in Chinese adults (Feng et al., 2011), and categorization tasks in Chinese and Israeli college students (Zhao & Bentin, 2008) among others (for reviews, see Hugenberg et al., 2010; Lee et al., 2011; Meissner & Brigham, 2001).

The own-race advantage has also been shown to extend to members of "intermediary" racial groups. For intance, multiracial participants have better memory for faces of racially ambiguous faces and are more likely to correctly categorize a face as multiracial than monoracial participants (Iankilevitch et al., 2020; Pauker & Ambady, 2009; Willadsen-Jensen & Ito, 2008). There is also some evidence of an analogous own-gender advantage, although the evidence for such an advantage is weaker, perhaps because people tend to have more comparable exposure to men and women than to their own versus other races (Herlitz & Lovén, 2013). However, to the best of our knowledge, past research has not investigated whether people with less common gender/sex identities and experiences (e.g., transgender, nonbinary individuals) differ from people without these identities and experiences in how they think about or respond to faces that vary in gender/sex.

Taken together, research on the own-race effect and to a lesser extent the own-gender effect, suggests that one's identity and the differences in social experience that come with it can affect how we think about, remember, and/or categorize faces. Therefore, to the extent that experiences align with commonly used social categories, people's judgment of the gender/sex or race in faces may also align with these social categories. Another, not mutually exclusive, possibility is that the categories themselves may play a more direct role in people's social judgments. Experiments assessing gender/sex and race face judgments have often used two categories (e.g., Campanella et al., 2001) and/or relied on stimuli that are predetermined to appear stereotypically male or female and monoracial in appearance (e.g., Chicago Face Database; Ma et al., 2015, London Face Database; DeBruine & Jones, 2021, Cafe Dataset; LoBue & Thrasher, 2015), with a few notable exceptions (e.g., Freeman, Rule, et al., 2010). Therefore, it is possible that instead of or in addition to experience shaping participants' use of binary categories to guide their judgments of sex and race, participants may use whatever categories come to mind or are made available to them within the framework and stimuli of a given face perception task. In line with this idea, people can come to use new categories within a single experimental session (Goldstone et al., 1996;

Zhou et al., 2010). Likewise, there is evidence that people will use a novel racial category to categorize faces that have been manipulated to appear racially ambiguous if provided with one (Chen & Hamilton, 2012; Chen et al., 2014).

Overview of the Current Studies

The current work investigated the extent to which people's judgments of social dimensions (gender/sex and race) align with discrete social categories (e.g., female, male, White, Black, etc.). We further explored some possible mechanisms that could explain the degree to which people's judgments align with discrete social categories.

Studies 1a and 1b built on past examinations of categorical judgments of gender/sex by asking participants to sort faces according to their gender/sex along a fully continuous scale, and by comparing how participants sorted faces to how they sorted nonsocial stimuli along nonsocial dimensions (color and number). The goal of these studies was to provide a stronger test of whether people sort faces into discrete categories (even when responding on a continuum) and whether this exceeds the extent to which participants exhibit similar behavior when making judgments about nonsocial stimuli. In Study 2, we investigated whether the pattern observed for gender/sex is also seen in another social dimension (race).

Studies 3–5 examined possible mechanisms through which judgments of faces along social dimensions come to align with social categories. First, Study 3 tested whether differences in people's identity and personal experiences—specifically whether they are cisgender or not—is linked to different response patterns when sorting faces along a gender/sex continuum. Then, we investigated whether cisgender people's sorting of intermediary faces can be modified through the introduction of a new category. We introduced participants to an intermediate third gender/sex (Studies 4 and 5) and race (Study 4), and asked if it changed where participants identified stimuli as belonging along a continuum. Finally, Study 5 tested whether the addition of a simple visual landmark at the center of the continuum had a similar effect on participants' sorting as the novel gender/sex category introduced in Study 4. Together, these studies provide a clearer picture of how people's judgments of social dimensions relate to discrete social categories as well as insight into the mechanisms that underlie these relations and the degree to which they are malleable.

Transparency, Openness, and Ethics

To ensure transparency and openness, all study materials, data, and code are available on the Open Science Framework project page (see https://osf.io/3xzq9/). Additionally, all resources developed by others are acknowledged and cited in the text and listed in the references section. All studies were preregistered (see https://osf.io/gb6tx/). Often the main article includes other/better analyses suggested by reviewers or discovered by the authors after a study was complete; however, for maximum transparency, the full set of preregistered analyses are in the online supplemental materials. Importantly, the findings highlighted in the main article reflect the conclusions of the parallel initially preregistered analyses. Participants provided informed consent before participating in our studies, which complied with ethical regulations and were approved by the Human Subjects Division Institutional Review Boards at the University of Washington and Princeton University. These preregistrations include all exclusion criteria used in the current article as well as additional analyses that go beyond the scope of the present article. However, all preregistered analyses are fully reported in the article or the online supplemental materials (see https://osf.io/d3htw/).

Studies 1a and 1b

In these studies, we investigated how people sort faces that range in masculinity and femininity along a continuous spectrum, compared to how they sort colors and numbers on a continuum, with a particular focus on how the most intermediary stimuli were sorted across the three domains. Participants saw a black line (i.e., continuum) with labels on either end, and were tasked with placing stimulus items (representing the domains of gender/sex, color, or number morphed at 10% increments; see examples in Figure 1 in the main article and see below for detailed information on stimuli for all studies) by clicking the location along the continuum where they thought that item was best represented in relation to the labels at either end.

Method

Participants and Exclusions by Study

In Study 1a, a convenience sample of 103 undergraduates from a public university in the United States completed the study for course credit in spring 2020. In all studies except Study 3 (see Study 3 Methods for more details), participants reported their gender identity using a multiple choice question with the options "female/woman," "male/man," "nonbinary," or "gender not listed here. Please specify [followed by a free-response textbox]." Additionally, participants in all studies reported their race using a multiple choice question where they checked one or more of the following options: "White/European," "Hispanic/Latino," "Black/African," "Asian," "Pacific Islander," "Native American/Alaska Native," or "other (please describe) [followed by a free-response textbox]." Finally, participants in all studies reported their age (in years) using an open-response textbox. We report demographic information for participants across all studies in Table 1. In all studies, we preregistered our sample sizes in advance of data collection, aiming for at least 100 subjects per cell. Here and in all following studies, we excluded data from trials when participants failed to give a response within the 4,000 ms response period or responded in 300 ms or under (a sign of careless responding), as per our preregistration (https://osf.io/u4dxw). In Study 1a, these time-based exclusions applied to 6.4% of the responses, including all data from one of the participants (not included in the total above, or in the demographics table).

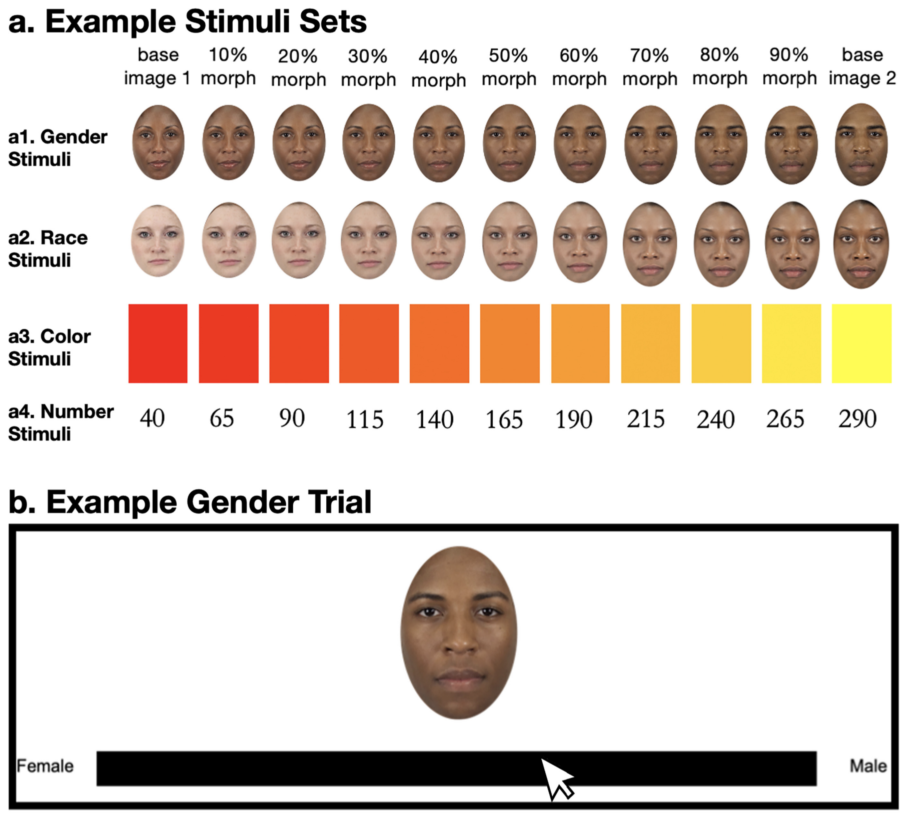
In Study 1b, a total of 99 individuals completed the study for monetary compensation on Prolific Academic in summer 2020. All participants had positive correlations between objective value and subjective judgment on the number of trials, so no participants were dropped based on that preregistered exclusion criterion. However, we dropped data from color trials of one participant who identified as color blind, as per our preregistration (https://osf.io/7b6qd). Finally, we excluded 3.4% of all responses based on response time.

Procedure

In Studies 1a and 1b, participants completed a sorting task via Inquisit Web (Millisecond Software, 2016) where they saw a black line (i.e., continuum) with labels at either end. On each trial, they were presented with a stimulus (in the domain of gender/sex, color, or number) and were asked to indicate the location where they believed

Figure 1

Example Stimuli and Example Gender/Sex Trial as Seen by Participants in Studies 1a, 1b, 2, and 3



Note. Panel A contains example stimuli in each domain presented throughout the article. The outermost stimuli were used as base images but were never seen by participants, while all nine inner morphs were presented to participants in a random order (along with other stimuli in the same category). In Studies 4–5, additional morphs at 5% increments were included for gender/sex and race stimuli. Panel B contains an example trial in the domain of gender/sex. From “The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data,” by D. S. Ma, J. Correll, and B. Wittenbrink, 2015, *Behavior Research Methods*, *47*, pp. 1122–1135. CC BY 4.0. See the online article for the color version of this figure.

it belonged along a continuum between the endpoints as quickly as they could (as in Figure 1, Panel B) based on their subjective judgment. For gender/sex trials, the target stimuli were morphed faces that were Asian, Black, or White (faces were morphed within racial group), and the labels were the words “female” and “male.” For color trials, the target stimuli were colors and the labels were the endpoint colors from which the target stimulus was created. Finally, for number trials, the target stimuli were numbers, and the labels were the endpoint numbers from the set to which the target stimulus belonged. In all three domains, stimuli varied regularly along the continuum at 10% increments (as demonstrated in Figure 1, Panel A; also see below for information about stimulus creation). Following completion of the sorting task, participants in both studies were redirected to a qualtrics survey where they completed a basic demographic questionnaire. In Study 1b only, this questionnaire also included the Gender/Sex Diversity Beliefs Scale (Schudson & van Anders, 2022), the Transphobia Scale (Nagoshi & Nagoshi, 2008), and a pair of feelings thermometers toward cisgender and transgender people, respectively (Alwin, 1997; Dunton & Fazio, 1997; IJzerman & Semin, 2009; see the online supplemental materials, pp. 4–5 for reporting on all survey measures).

Task Design

All participants completed three blocks of 54 gender/sex trials each, interspersed with the color and number blocks (which were counterbalanced across participants and also contained 54 trials each) between them. In total, participants sorted 270 stimuli exactly once, and all stimuli were presented in a random order within each category domain. In both Studies 1a and 1b, labels were stationary (e.g., “female” was always on the left), though this varied in later studies. In Study 1a, stimuli were completely randomized within category across all blocks, but in Study 1b, gender/sex stimuli were grouped based on the base pair used to generate them (e.g., all nine faces generated by a single male/female pair were always run in the same block) to ensure our observed results were not design-dependent.

Task Stimuli

Gender/sex stimuli for Studies 1a and 1b were created by randomly pairing the 12 most masculine male faces and 12 most feminine female

Table 1
Demographic Information for All Studies

Study number	Study 1a	Study 1b	Study 2	Study 3	Study 4	Study 5
Current gender/sex						
Woman/female	61	42	99	117	83	126
Man/male	41	54	46	139	141	191
Nonbinary	0	3	3	72	4	6
Agender	0	0	0	10	0	0
Genderfluid	0	0	0	11	0	0
Genderqueer	0	0	0	9	0	0
Another gender category	0	0	0	33	0	0
Did not self-identify	1	0	0	0	0	0
Race						
Asian	50	11	44	14	20	15
Black	5	2	25	29	10	3
Latinx	6	3	0	53	7	44
White	30	80	65	258	180	244
Multiracial	8	2	14	32	10	13
Another racial category	3	1	0	5	1	4
Age						
M (in years)	19.64	26.48	19.36	25.14	28.97	25.11
SD (in years)	1.37	7.64	1.20	6.87	10.92	8.02
Total *N* (per study)	103	99	148	391	228	323

faces in each of three races (Asian, Black, and White) from the Chicago Face Database (Ma et al., 2015). The resulting 36 male–female, monoracial face pairs served as the endpoints, which were systematically morphed using the WebMorph tool (DeBruine, 2018) at a 10% gradient to create nine 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 for an example set of face gender/sex stimuli and pp. 1–2 in the online supplemental materials for additional details about our face morphing procedure for all face stimuli in this article). 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/ 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 greater 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 six distinct pairs of endpoints. Each of the six colors was represented twice to create six total continua, and we purposefully avoided pairing complementary colors due to their propensity to cancel each other out (or lose hue) and produce a grayscale color when mixed. Then, each pair was systematically morphed from at a 10% gradient, resulting in nine intermediary colors that were equidistant from one another in RGB color space for each of the six pairs (see Figure 1 for an example set of color stimuli). At the test, participants sorted all 54 color stimuli in both Studies 1a and 1b (see https://osf.io/u69yj/ for a full set of color stimuli).

Number stimuli for Studies 1a and 1b were created by using a multistep process. First, we used https://random.org to generate six distinct multiples of 10 between 0 and 100 to serve as lower values. Then we generated six 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 nine intermediary values that were equidistant from one another for each of the six 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/ for a full set of number stimuli).

Task Scoring

In Studies 1a and 1b, our two variables of interest were subjective judgment (i.e., the place on the bar that people clicked to sort each stimulus relative to the endpoints), and latency (i.e., how long it took for participants to sort the target stimulus; all latency results are reported in the online supplemental materials on pages 2 and 5). 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 and 100 proportional to the place along the line that was clicked. In both studies, we computed participants' per-trial judgment error as 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. If a participant's subjective judgment was at 75%, the judgment error would be 5. Note that "error" here only refers to the distance between participants' judgments and the target location according to the morph calculation. Computing participants' error enabled us to compare participants' estimates at different points along the continuum and between conditions. However, it should not be confused with inaccuracy in a broader sense, such as a reflection of how close participants estimates' were to some "true" gender/sex or racial variation in the world. We return to this issue in the General Discussion.

Overview of Statistical Approach

In the six studies presented in this article, we used two types of regression models to investigate how participants sorted stimuli on the continuum in the domains of color (Studies 1a and 1b), number (Studies 1a and 1b), race (Studies 2–4), and gender/sex (all studies). In all studies, we examined participants' error when sorting stimuli within and across each domain and investigated whether responses to the more intermediate stimuli (i.e., those with an objective location closer to the center of the continuum) were placed just as precisely as stimuli that fall closer to a category boundary, or if people show more uncertainty in sorting intermediary stimuli. To do this, we ran multilevel regression models using the "lmer" function in the lme4 package in R (Bates et al., 2014) predicting error from intermediateness (henceforth used to describe the distance from the endpoints of the scale), and other factors that we specify below in the results section of each study. In all our models, we also included a random intercept for participant ID, as well as random slopes within-participant for intermediateness. In studies where we investigated a second within-subjects predictor variable (e.g., stimulus domain in Studies 1a, 1b, 2, and 4, and response window, in Study 3), we included a random slope within-participant for this variable in our model[4] and set a reference group for each model (see individual results sections for the reference group set in each study). Thus, all regression results in our main article are presented within the context of the reference group (e.g., we report the effect of ambiguity within gender as opposed to a main effect of ambiguity). Degrees of freedom for all models were calculated using Satterthwaite's method, which takes into account the complex random effects structure in the model (Kuznetsova et al., 2017). To further understand patterns of accuracy across stimuli, we visually inspect and describe sorting patterns at different levels of stimulus ambiguity across all three domains.

Note that in Studies 1a, 1b, 2, 3, and 4, we also ran similar multilevel regression models (also using lmer) to investigate how participants' performance on the gender/sex version of the task differed by the race of the faces used in the stimuli. Although we found some differences by face race in gender task performance, the pattern of findings across studies was not consistent and therefore further discussion of these analyses is contained in the online supplemental materials, see pp. 3–4 (Study 1a), p. 5 (Study 1b), p. 8 (Study 2), p. 22 (Study 3), and p. 31 (Study 4). Finally, in Studies 1a, 1b, and 2, we looked to see if the overall relationship between participants' responses and the objective location of the corresponding morphs along the continuum aligned with findings from past work in the domains of color (Winawer et al., 2007), number (Feigenson et al., 2004), and gender/sex (Campanella et al., 2001; Freeman, Rule, et al., 2010). To do this, we used the "geeglm" function in the geepack package (Højsgaard et al., 2006) to run generalized linear model polynomial regression models, to regress participant's subjective judgments onto linear, quadratic, cubic, and quartic components of objective location within each domain, clustering on participant ID. In our main article, we report analyses run in R Studio (Version 2023.03.1+446) and use scaled-down values of error, intermediateness, subjective judgments, and objective location, as applicable for each model.

Results

To examine the effect of intermediateness (distance from the nearest endpoint) on error in the three domains, we ran multilevel regression models predicting error from domain, intermediateness, and their interaction. In these models, we included a random intercept for participant ID as well as random slopes within-participant for intermediateness and for domain and set gender as the reference group. We found that across domains, intermediateness predicted greater error ($b = .110$, $t[104] = 10.45$, $p < .001$ in Study 1a; and $b = .110$, $t[110] = 15.72$, $p < .001$ in Study 1b). Moreover, the influence of intermediateness on error was more pronounced for gender/sex relative to the other two domains, meaning that people were less accurate at sorting intermediary gender/sex relative to color ($b = -.042$, $t[25400] = -6.99$, $p < .001$ in Study 1a; and $b = -.029$, $t[25900] = -5.15$, $p < .001$ in Study 1b) and relative to number ($b = -.100$, $t[25400] = -16.41$, $p < .001$ in Study 1a; and $b = -.084$, $t[26000] = -14.88$, $p < .001$ in Study 1b). Thus for gender/sex in particular, people were especially inaccurate at estimating the location of stimuli for the most intermediary/androgynous faces (see Figure 2).[5]

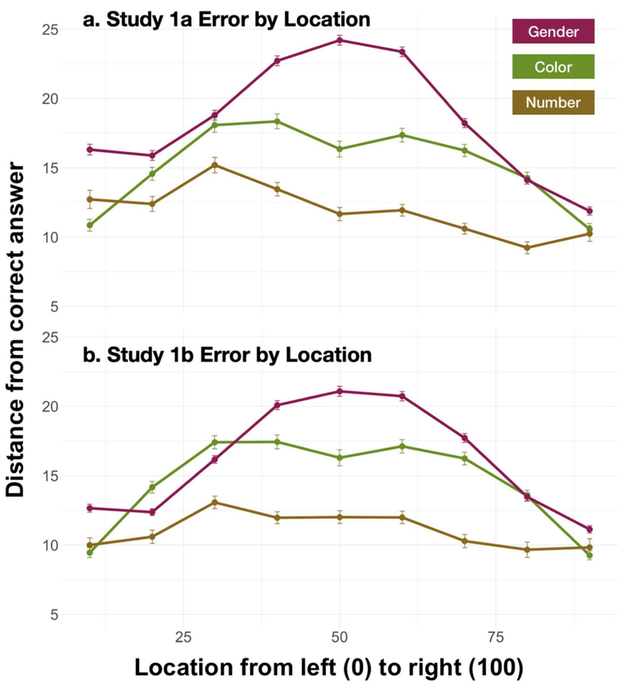
In Figure 3, we show the range of responses from participants at each objective stimulus location. Responses to the numeric stimuli appear fairly linear and as can be seen in Figure 2, there are fairly equal amounts of error at each objective location. In contrast, responses to the color stimuli appear trimodal (Figure 3) with increasing levels of error between the endpoints and midpoint. Finally, responses to face stimuli are bimodal for the more prototypical stimuli, but for the more intermediary stimuli (50% morphs), participants' judgments were distributed across the entire range of possible values. Further, their error in judgments was highest for these intermediate stimuli (Figure 2).

In addition to our visual inspection, we found evidence of both linear and a nonlinear relation between participants' responses and the objective location of the corresponding morphs along the continuum across all domains. Using a generalized linear polynomial regression model, we regressed subjective gender/sex judgments onto linear, quadratic, cubic, and quartic components of objective location within each domain. For gender/sex, we found that in addition to a linear relationship between location and subjective judgment (Study 1a: $b = .118$, $SE = .004$, $p < .001$; Study 1b: $b = .128$, $SE = .003$, $p < .001$), participants judgments indicated significant cubic (Study 1a: $b = -.003$, $SE = .0002$, $p < .001$; Study 1b: $b = -.003$, $SE = .0001$, $p < .001$) components, suggesting a nonlinear relationship between objective and subjective gender/sex judgments (see Figure 2). Similarly, for numbers, participants' judgments indicated significant linear (Study 1a: $b = .008$, $SE = .003$, $p < .001$; Study 1b: $b = .086$, $SE = .002$, $p < .001$), and quadratic (Study 1a only: $b = -.004$, $SE = .001$, $p = .002$) components. For colors, participants' judgments suggested a significant linear component (Study 1a: $b = .103$, $SE = .003$, $p < .001$; Study 1b: $b = .114$, $SE = .003$, $p < .001$) as well as a cubic component (Study 1a: $b = -.001$, $SE = .0002$, $p < .001$; Study 1b: $b = -.001$, $SE = .0002$, $p < .001$). Collectively, these results suggest that there both linear and nonlinear components of the relationship between objective and subjective judgments of gender, color, and number, and that the linear component was consistently strongest across all domains.

[4] We attempted to run models that also included a random intercept for face-pair, but found that some of these models failed to converge. Thus, for the sake of consistency across models, we opted to report all models without an additional intercept for face-pair.

[5] For Studies 1a, 1b, 2, 3, and 4, we also used similar multilevel regression models to investigate how participants' performance on the gender/sex version of the task differed by the race of the face stimuli. We found some evidence that there was greater perceptual error for Black and Asian faces with intermediary gender/sex, relative to that of White faces with intermediary gender/sex. Further discussion of these analyses is contained in the online supplemental materials.

Figure 2
Average Distance From Correct Answer (or Error) by Objective (Correct) Location and by Domain in Studies 1a and 1b



Note. Error bars represent standard error. See the online article for the color version of this figure.

Discussion

Across two studies that varied several aspects of study design, we consistently found evidence of both linear and nonlinear relationships between participants' judgments of where a stimulus belonged on a continuum and its objective location (based on morph level) across all domains. Looking more closely, we found that faces closer to the endpoints were sorted bimodally, while the sorting of intermediary (or androgynous) faces spanned the entire length of the continuum, resulting in greater error. This pattern replicated across two participant samples, sets of stimuli, and task structures.

These findings reinforce previous work suggesting that participants sort gender/sex in a binary way but lack consensus about the location of intermediary faces (Campanella et al., 2001; Wild et al., 2000). Notably, this pattern is distinct from participants' sorting of numbers and colors but still leaves us uncertain as to whether the sorting patterns we found in Studies 1a and 1b are unique to gender/sex or also generalize to other types of face perception. We test this in our next study, where we compare patterns of sorting behavior between gender/sex and race.

Study 2

In Study 2, we tested whether a pattern of responses similar to the gender/sex sorting in Studies 1a and 1b would characterize participants' judgments of race. We specifically chose race as a comparison group because, like gender/sex, race is often treated as categorical (Freeman, Pauker, et al., 2010; Gaither, 2015) even though it exists along a continuum. We measured people's sorting patterns using a nearly identical task to the ones used in Studies 1a and 1b, but instead of asking people to sort colors or numbers, participants sorted faces by gender/sex and race. As in Studies 1a and 1b, we also examined the error reflected in the distribution of participants' sorting within and across the domains of both gender/sex and race.

Method

Participants and Exclusion

In Study 2, a convenience sample of 148 undergraduates from a private university in the United States completed the study for course credit in fall 2020. We dropped 2.37% of task responses, including all data from one additional participant (not included in the total above) based on response time criteria outlined in Study 1a. Finally, we found that all participants showed positive correlations between subjective judgment and objective location for race so no participants were dropped as a result of that preregistered exclusion criterion.

Procedure

Study 2 was very similar to Studies 1a and 1b, except that in this iteration of the task, participants sorted two blocks of race stimuli (which contained equal numbers of faces that were male and female) and two blocks of gender/sex stimuli (which contained equal numbers of faces that were Asian, Black, or White) in random order each containing 54 trials (see Figure 1, Panel A, rows A1 and A2). As in prior studies, participants were asked to complete a short survey following the main 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, 2008), and basic demographic information (again, see the online supplemental materials, pp. 7–8 for reporting on survey measures).

Task Design

As in Study 1b, stimuli were grouped together based on the base pair of faces used to create them. Additionally, we counterbalanced the location of labels (which were male and female for gender/sex 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 we collected and scored the same variables of interest (latency and subjective judgment) as in Studies 1a and 1b.

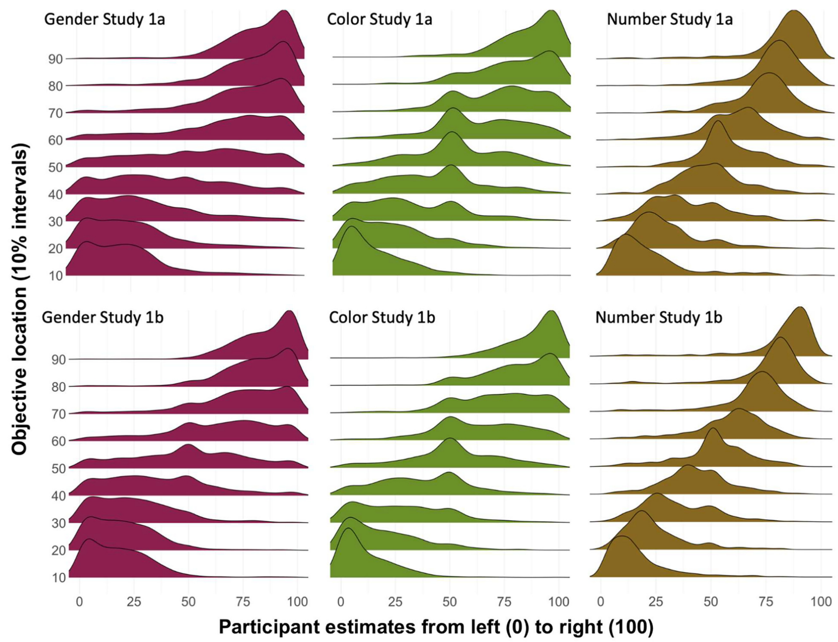
Task Stimuli

Gender/sex 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, 2021) to enhance stimulus sampling and generalize our findings beyond faces in the Chicago Face Database (Wells & Windschitl, 1999). The resulting set of gender/sex stimuli consisted of four same-race face pairs from each of three racial groups (Black, East Asian,[6] and White) morphed at a 10% gradient for a total of 108 faces that participants sorted at the test.

[6] As opposed to West Asian, as denoted in the London Face Database.

Figure 3

Distribution of Participant Estimates of Stimulus Location Along the Continuum by Objective (Correct) Location in Studies 1a and 1b



Note. Error bars represent standard error. See the online article for the color version of this figure.

Race stimuli were created by selecting the four male faces and four 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 three races) from the Chicago Face Database (Ma et al., 2015). Then, we semirandomly paired faces across race (i.e., Asian-White, Black-Asian, and White-Black) to end up with two male and two female pairs for each cross-race combination. The resulting 12 face pairs served as the endpoints, which were systematically morphed at a 10% gradient to create nine distinct morphs (see Figure 1 for an example set of race stimuli). At the test, participants sorted all morphs (but not original faces) for a total of 108 faces (see https://osf.io/u69yj/ for a full set of race stimuli).

Results

Using the same modeling approach as in Studies 1a and 1b (i.e., a multilevel regression predicting error from domain, intermediateness, and their interaction, along with a random intercept for participant ID, random slopes within-participant for intermediateness and for domain, and gender set as the reference group), we replicated the intermediateness effect from Studies 1a and 1b showing again that participants demonstrated greater error when judging more intermediate stimuli overall ($b = .101$, $t[161] = 10.58$, $p < .001$), and in each domain ($b = .090$, $t[147] = 10.20$, $p < .001$ for race, and $b = .100$, $t(146) = 8.79$, $p < .001$ for gender/sex). Moreover, the influence of objective location on error was slightly stronger for gender/sex relative to race, such that people were slightly less accurate at judging intermediary gender/sex relative to race ($b = -.011$, $t[31000] = -2.67$, $p = .008$; also see Figure 4, Panel B, for visualization). Furthermore, participants were less accurate at judging gender/sex relative to race ($b = -.021$, $t[29100] = -6.32$, $p < .001$). Again, participants' responses for intermediate stimuli were distributed across the full range of possible responses, though visually there appeared to be more of a peak for race than for gender/sex, reflecting the greater error in judgments of race (see Figure 4, Panel A).

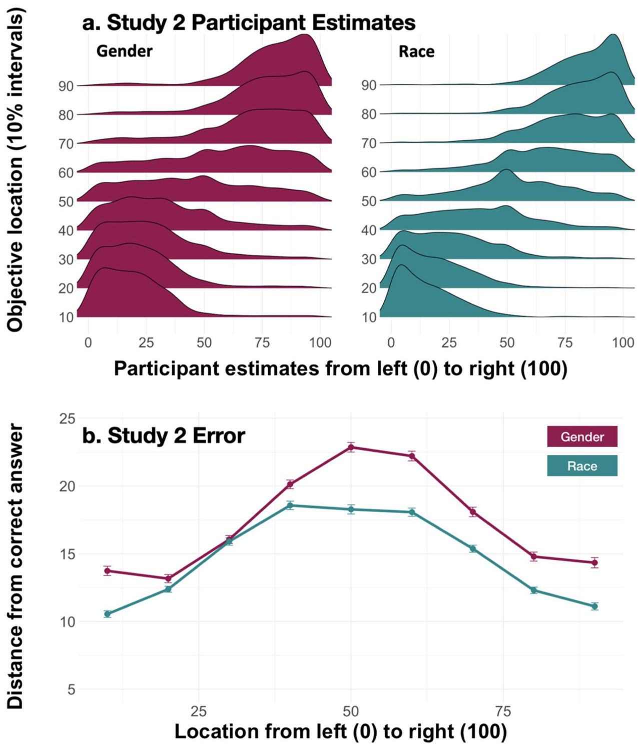
As before, we also used generalized linear model polynomial regression, to examine the relationship between subjective judgment and objective location within each domain. We found that this relationship contained linear ($b = .114$, $SE = .003$, $p < .001$), quadratic ($b = .008$, $SE = .001$, $p < .001$), cubic ($b = -.003$, $SE = .0001$, $p < .001$), and quartic ($b = -.0003$, $SE = .0001$, $p < .001$) components for gender/sex, and race (linear: $b = .131$, $SE = .003$, $p < .001$; quadratic: $b = -.005$, $SE = .0009$, $p < .001$; cubic: $b = -.003$, $SE = .0001$, $p < .001$; quartic: $b = .0002$, $SE = .00005$, $p < .001$) suggesting a strong linear relationship and some evidence of nonlinear relationships between objective location and subjective judgment in both gender/sex and race judgments (see Figure 4, Panel A).

Discussion

In Study 2, we replicated the gender/sex results from Studies 1a and 1b and found a similar (though slightly weaker) pattern in the domain of race. Specifically, we found that as in the domain of gender/sex, there was evidence of both linear and nonlinear relationships between objective location and subjective judgment of race, and that people exhibited greater error and variability when sorting intermediary stimuli than those stimuli that were objectively closer to the endpoints. Together, Studies 1 and 2 provide substantial evidence that participants display less consensus about the location of intermediary faces and are less accurate when sorting more intermediary faces.

Figure 4

Distribution of Participant Estimates of Stimulus Location Along the Continuum (Panel A) and Average Distance From Correct Answer (Panel B) in Study 2



Note. Panel A is the distribution of participant estimates of stimulus location by objective location in Study 2 and Panel B is average distance from correct answer (or error) by objective (correct) location and by domain in Study 2. Error bars represent standard error. See the online article for the color version of this figure.

Study 3

Study 3 investigates the possible role of experience in participants' judgments of faces. We specifically ask if gender-diverse people—who have more experience with gender/sex outside of the traditional binary of "(cisgender) man" and "(cisgender) woman"—show more accuracy, particularly in sorting more androgynous (or intermediary) faces. There are several not mutually exclusive reasons why gender-diverse people (defined as transgender, nonbinary, and genderqueer) might show greater accuracy in sorting more androgynous faces. One possibility is that gender-diverse people have a more expansive and less binary view of gender/sex in general (Iantaffi & Bockting, 2011; Nagoshi et al., 2012). This may occur as a result of their own lived experiences with gender transition (deviating from the default gender identity associated with their assigned sex; Bradford et al., 2019; Vries, 2012) or identifying as a gender that is not binary. In addition to increased conceptual flexibility in thinking about the gender/sex category, it is also possible that gender-diverse participants have greater perceptual fluency for androgynous faces as a result of self-selecting into social groups with a higher proportion of people whose appearance physically changed across this spectrum via medical transition (e.g., through hormones and/or surgeries; Factor & Rothblum, 2008; Gagné et al., 1997; Galupo et al., 2021), thus resulting in greater exposure to androgyny than our cisgender sample. Gender-diverse people may also have attended more to gender, having noted how they and others are treated as a function of their gender/sex and appearance (Levitt & Ippolito, 2014). A fourth possibility is that gender-diverse participants are more motivated to avoid misgendering people (Richards et al., 2016) which could impact their performance on the task as well as how they conceptualize gender/sex more generally. According to any of these possibilities or their combinations, we would expect gender-diverse participants to be more accurate than cisgender heterosexual participants at sorting faces along a continuous scale, particularly for faces that are more androgynous.

In this study, we also manipulated one other aspect of the design that differs from our previous studies. Specifically, we had participants complete the task within a tighter time constraint (within 2,000 ms[7]) and at the same speed as in the previous studies (4,000 ms). This manipulation was selected in anticipation of a group difference between the gender-diverse and cisgender heterosexual participants at 4,000 ms. We anticipated that if we observed that difference, one potential explanation could be that the gender-diverse group was consciously, deliberately selecting choices in the center so as not to be biased by the more binary categories. We reasoned that if this was a deliberative strategy, it would be harder to deploy when given less time. Alternatively, if the gender-diverse sample's tendency to think of gender/sex as more continuous was more of an automatic process, or if it was a perceptual fluency effect, we would expect the group difference to also be observed at the faster time limit as well as the slower one.

Method

Participants and Exclusions

In Study 3, a total of 391 participants who self-identified as either gender diverse ($n = 194$) or cisgender heterosexual ($n = 197$) completed the study for monetary compensation on Prolific Academic in January 2022. Unlike other studies in this article, participants in Study 3 reported both their gender identity and sexual orientation using (two separate) open-response text boxes. For a comprehensive report of participants' responses, please see Table S1 in the online supplemental materials (pp. 10–11). Similar to prior studies, we excluded data from trials where participants responded in 300 ms or less and data from participants who indicated face blindness. In this study, we also preregistered to exclude anyone who exceeded the maximum amount of time allotted to sort each face (2,000 ms in the "fast" condition and 4,000 ms in the "slow" condition; see task design for more information on the time manipulation) for more than 10% of the trials. In sum, these criteria resulted in the exclusion of 8.77% of all responses and full data from nine participants (not included in the total above).

Procedure

Study 3 used the same continuous sorting task used in previous studies, but unlike previous iterations, participants in this study only sorted faces on the basis of gender/sex. Additionally, we programmed the task such that participants had either 2,000 ms (in the "fast"

[7] Chosen based on the results of a pilot test run during summer of 2021; for more information see the online supplemental materials, p. 9.

condition) or 4,000 ms (in the "slow" condition) to make each judgment. To maintain an entirely within-subjects design, all participants were asked to sort one block of stimuli "fast" and one block of stimuli "slow" (order was counterbalanced across participants, and each block contained equal numbers of faces that were Asian, Black, or White). Also like in prior studies, participants were asked to fill out a short survey following completion of the task which included questions about their experience completing the task, the Gender Identity Reflection and Rumination Scale (Bauerband & Galupo, 2014), the Genderqueer Identity—Theoretical Awareness Scale (McGuire et al., 2019), and general demographic information (again, see Table 1).

Task Design

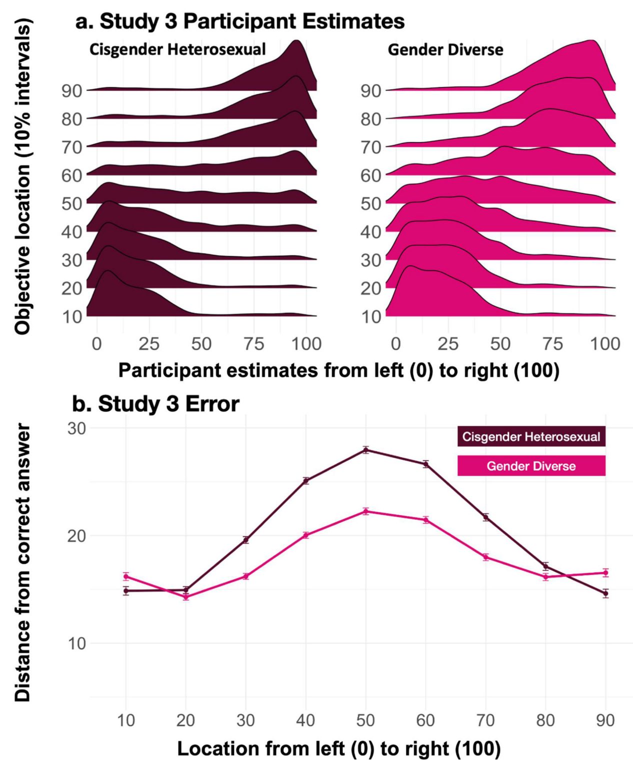
Each block contained six same-race face pairs from each of three racial groups (Black, East Asian, and White) morphed at a 10% gradient for a total of 108 faces that participants sorted at the test. As in previous studies, stimuli were grouped together based on the base face pair used to create them and were presented in a random order within their respective morphing set. Additionally, we counterbalanced the location of labels (which were male and female for gender/sex trials and Asian, Black, or White for race trials) so that participants did not associate a label with one particular side of the continuum. As before, data were collected using Inquisit Web (Millisecond Software, 2016) and we collected and scored the same variables of interest (latency and subjective judgment) as in Studies 1 and 2.

Task Stimuli

Stimuli for Study 3 consisted of a subset of the stimuli used in Study 1a. These 108 faces were created by randomly pairing four of the most phenotypically masculine and four of the most phenotypically feminine faces within Asian, Black, and White races from the Chicago Face Database (Ma et al., 2015) morphed at a 10% gradient.

Figure 5

Distribution of Participant Estimates of Stimulus Location Along the Continuum (Panel A) and Average Distance From Correct Answer (Panel B) in Study 3



Note. Panel A is the distribution of participant estimates of stimulus location by objective location in Study 3 and Panel B is average distance from correct answer by objective location and by domain in Study 3. Error bars represent standard error. See the online article for the color version of this figure.

Results

We ran a multilevel regression model predicting error from intermediateness, participant group, response window, and their two and three-way interactions (see the online supplemental materials, p. 11 for full model) and found that again, all participants (regardless of gender/sex identity) demonstrated overall greater error when sorting more androgynous stimuli (i.e., stimuli that were more intermediary in the domain of gender/sex) than more gendered stimuli (i.e. gender/sex stimuli that were closer to the endpoints; $b = .150$, $t[458] = 13.06$, $p < .001$). However, this effect was attenuated for gender-diverse participants in comparison to the cisgender heterosexual participants such that gender-diverse participants demonstrated less perceptual error than cisgender heterosexual participants at sorting more androgynous stimuli ($b = -.082$, $t[451] = -5.07$, $p < .001$) which supports the possibility that gender-diverse participants may have more conceptual flexibility in the domain of gender/sex than cisgender heterosexual participants (see Figure 5).[8]

Follow-up analyses investigating group differences in error at each location along the continuum revealed that there were no significant differences between cisgender heterosexual participants and gender-diverse participants when sorting less androgynous faces (i.e., faces that were closer to the endpoints of the continuum; all $ps > .1$; see Figure 5; Table S2 in the online supplemental materials). However, gender-diverse participants demonstrated significantly less error at categorizing more androgynous faces (i.e., faces that were closer to the middle of the continuum) than their cisgender heterosexual counterparts ($b = -3.54$, $p < .001$, for faces located 20 away from the middle of the continuum; $b = -5.46$, $p < .001$, for faces located 10 away from the middle of the continuum; $b = -5.70$, $p < .001$, for faces located at the middle of the continuum; also see Figure 5; Table S2 in the online supplemental materials). Unlike androgyny, we found no meaningful differences in the error of participants' judgments based on the amount of time they were given (2,000 vs.

[8] In the preregistration, our two primary groups of interest were cisgender heterosexuals and gender diverse participants. However, we also preregistered several secondary exploratory analyses involving a third group of cisgender participants who are queer. Additional information and full results from the cisgender-LGBPQ+ group can be found in the online supplemental materials starting on page 12. Overall, the pattern of responses for this group generally fell between the two primary groups of interest, though often closer to the cisgender heterosexual group.

4,000 ms) to sort each face (see the online supplemental materials, pp. 14–15 for more extensive discussion).

Discussion

We found that gender-diverse participants judged the gender/sex of faces more accurately than the cisgender heterosexual participants and that this effect was driven by gender-diverse participants' reduced error in their perception of more androgynous faces relative to their cisgender heterosexual counterparts. We observed the group difference at both faster and slower response windows.

Whether the group difference was driven by the gender-diverse sample's conceptual understanding of gender/sex as more continuous, particular first-hand perceptual experience with androgyny, the salience of gender, gender transitions, or differences in motivation cannot be determined by this study. However, the following studies aim to ask whether people who lack the long-term experiences of gender-diverse people might still be capable of a change in categorization performance through a comparatively brief experimental manipulation.

Study 4

In our next studies, we sought to further investigate the influence of experience on gender/sex and race judgments via experimental manipulation. Specifically, we wanted to follow-up on the idea that a change in one's knowledge or attention to particular gender categories could influence one's judgments of gender/sex. To do so, we recruited a new sample of cisgender participants and tested whether or not exposure to a third category at the center of the scale during the task would improve their judgment of intermediary faces.

Method

Participants and Exclusions

In Study 4, a total of 228 individuals (randomly assigned to either the two-category or three-category conditions; $n = 114$ in both conditions) completed the study for monetary compensation on Prolific Academic in winter 2021. We excluded 8.28% of task responses, including all data from eight participants (not included in the total above) based on response time criteria outlined in Study 1a. In line with our preregistration, we also excluded data from participants whose correlations between subjective judgment and objective location were not positive within a given block, resulting in the exclusion of an additional 5.34% of the remaining task responses, including all data (i.e., both blocks) from an additional four participants (not included in the total above).

Procedure

Participants in Study 4 were randomly assigned to complete either a two-category label (similar to past studies) or three-category label version of the task. In both conditions, participants were introduced to face categories represented by symbols (e.g., sigma, omega). Before completing the task, participants were told they were being asked to sort people associated with different symbols. Before each block, participants in the two-category condition were taught to associate two sets of images with two symbols, each intending to represent a distinct race or gender/sex category (see Figure 6, Panel A, outer images). This approach was adopted because there was not an obvious verbal label for the intermmediate cases; therefore all labels were converted to symbols. Participants in the three-category condition also saw a third set of images and an additional label, which represented an intermediate category that was presented to participants along with two 50% morph faces (see Figure 6, Panel A center). 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 in the online supplemental materials, 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 in the online supplemental materials, Panel C).[9] On test trials, participants saw the continuum with all the labels (as symbols) they learned about placed at the expected locations (including in the middle, if in the three-category condition), and were asked to sort each stimulus along the continuum (see Figure 6, Panel B for example test trials). In total, participants completed two blocks of 76 trials in random order, one containing race stimuli and the other containing gender/sex stimuli.

Task Design

As before, stimuli in Study 4 were grouped together based on the base face pair used to create them, and we counterbalanced the location of labels (i.e., symbols) on either end 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 the completion of the task which included basic demographic information.

Task Stimuli

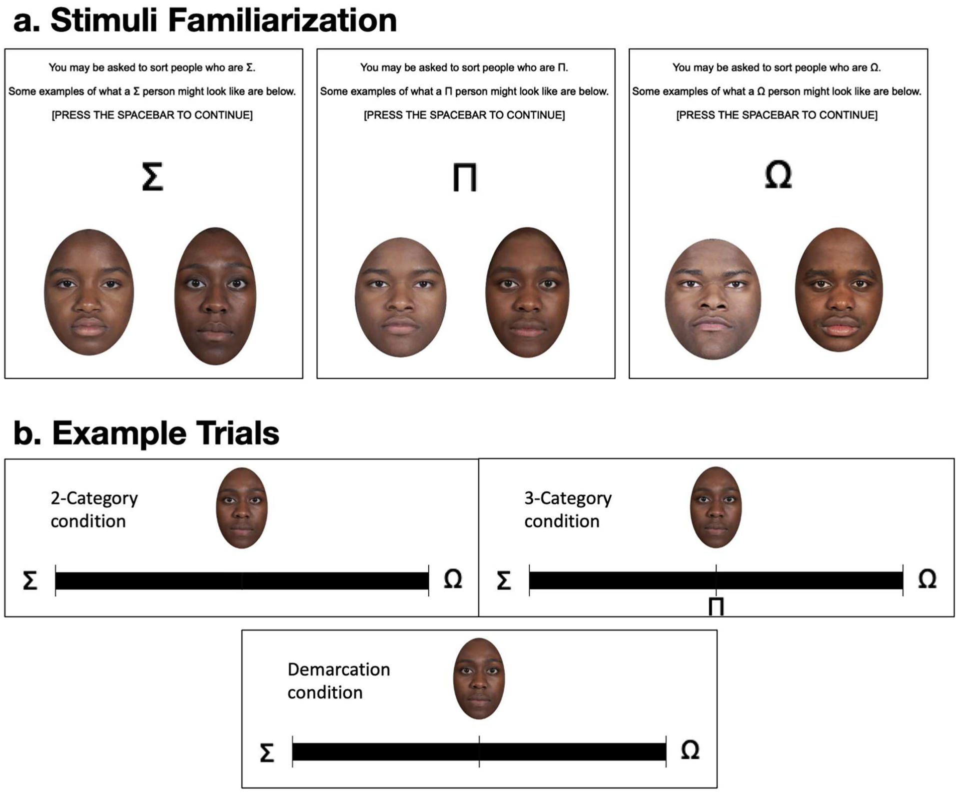
Both gender/sex and race stimuli used in Study 4 were created from the same set of base faces used in Study 2. Unlike in previous studies, each of the 12 base pairs was morphed at a 5% gradient to create a finer gradation of 19 distinct intermediate morphs resulting in a set of 228 faces for both race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

Results

As in previous studies, we conducted a multilevel regression model predicting error from intermediateness, condition, domain, and their two- and three-way interactions. We also included a random intercept for participant ID and random slopes within-participant for intermediateness and for domain. Here we set our reference groups to compare gender to race and the two-category condition to the three-category condition. We found that participants demonstrated greater error when perceiving more intermediary stimuli ($b = .100$, $t[265] = 8.34$, $p < .001$; see Figure 7). Participants in the three-category condition demonstrated greater overall error than participants in the two-

[9] Although participants did not complete practice (or test) trials prior to completing the study, we removed participants who clearly did not understand the task by excluding people whose judgments did not positively correlate with the objective location of each stimulus (which was only possible if they learned the association), which resulted in removal of only four participants' data. We also solicited optional open response feedback about the task and participants did not report confusion about the task.

Figure 6
Example Stimuli Familiarization for Three-Category Condition and Example Task Trials as Seen by Participants in Studies 4 and/or 5



Note. Panel A: Example stimuli familiarization in the three-category gender/sex condition; Panel B: Example 2 and three-category task trials (as seen in Studies 4 and 5) and demarcation task trial (as seen in Study 5 only). See the online article for the color version of this figure.

category condition ($b = .032$, $t[238] = 3.24$, $p = .001$). Critically, participants in the three-category condition showed less error for intermediate faces than those in the two-category condition ($b = -.074$, $t[273] = -4.34$, $p < .001$), suggesting the addition of the third category reduced judgment error for the intermediary morphs.

To make sure participants in the three-category condition were not merely choosing locations closer to the midpoint regardless of the objective location (potentially causing the improved performance around the midpoint and worsened performance at the ends), we ran a second multilevel regression model predicting subjective distance from the midpoint (computed as the absolute distance between participants' subjective judgments and the midpoint of the continuum), with domain, condition, intermediateness, and their interactions as predictors, along with a random intercept for participant ID and random slopes within-participant for intermediateness and for domain. Results revealed that less objective distance from the midpoint was associated with less subjective distance from the midpoint ($b = -.100$, $t[316] = -11.40$, $p < .001$). This was less true for participants in the two-category condition than those in the three-category condition ($b = -.030$, $t[329] = -2.34$, $p = .020$), and less true in the domain of gender/sex than race ($b = -.066$, $t[309] = -9.31$, $p < .001$).

In addition to the formal statistical analysis, visual inspection of the results in Figure 8 makes clear that participants showed bimodal responses for the more prototypical stimuli in the two-category conditions while showing a more trimodal response for the three-category. This finding suggests people were able to adopt a third category.

Discussion

Participants assigned to think of the line as containing members of three categories (vs. the default two endpoints only), were more accurate in sorting more intermediary faces. These results suggest that cisgender participants were able to adopt a third category into their conceptualizations of both gender/sex and race with minimal exposure (i.e., two exemplars). Although the inclusion of a third category at the midpoint increased overall error, likely due to the increased load of remembering all three categories, it decreased participants' judgment error for intermediary faces (its target phenotype). However, an alternative interpretation, that we address in the final study, is that by marking the center, we created a salient landmark between the two existing categories, rather than engendering a third category concept.

Study 5

In Study 5, we distinguish two possible interpretations of Study 4—that the three-category condition increased the saliency of the center of the continuum or that it created an actual third category

Figure 7
Average Distance From Correct Answer by Objective Location and by Domain and Condition in Studies 4 and 5

Note. Error bars represent standard error. See the online article for the color version of this figure.

in the minds of our participants. In Study 5, we compare three conditions, the two- and three-category conditions from Study 4, and a new "simple demarcation" condition in which the center of the line is bifurcated, calling attention to it (henceforth, demarcation condition). If our findings in Study 4 are a result of saliency, we would expect that participants in Study 5 would perform equally well in the three-category and demarcation conditions. However, if there is an additional advantage to the adoption of a new category above and beyond a saliency effect, we expect that participants in the three-category condition will perform better in sorting more androgynous stimuli than those in the demarcation condition.

Method

Participants and Exclusions

In Study 5, a total of 323 individuals (randomly assigned to either the two-category condition; $n = 112$, three-category condition; $n = 104$, or demarcation condition; $n = 107$) completed the study for monetary compensation on Prolific Academic in spring 2021. We excluded data from 5.3% of task responses, including all data from 15 participants (not included in the total above) based on response time criteria outlined in Study 1a. In line with our preregistration, we also excluded data from participants whose correlations between subjective judgment and objective location were not positive, resulting in the exclusion of an additional 6.4% of the remaining task responses, including all data from an additional 22 participants (not included in the total above).

Procedure

The procedure for Study 5 was similar to that of Study 4 with one key difference. In addition to the two-category label and three-category label version of the task, we introduced a third condition that was identical to the two-category condition except that participants in this condition sorted stimuli on a black line with a demarcation (but no label) at the midpoint (i.e., demarcation condition). Participants then went through the same training procedure as Study 4 where they 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 two-category condition), and then saw two example faces sorted at 25% and 75% of the way between the endpoints of the continuum. At the test, participants saw a black line with two labels (two-category condition), three labels (three-category condition), or two 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 total, participants in Study 5 completed just a single block of trials, sorting along the dimension of gender/sex. Finally, participants were asked to fill out a short demographic questionnaire following the completion of the task.

Task Design and Stimuli

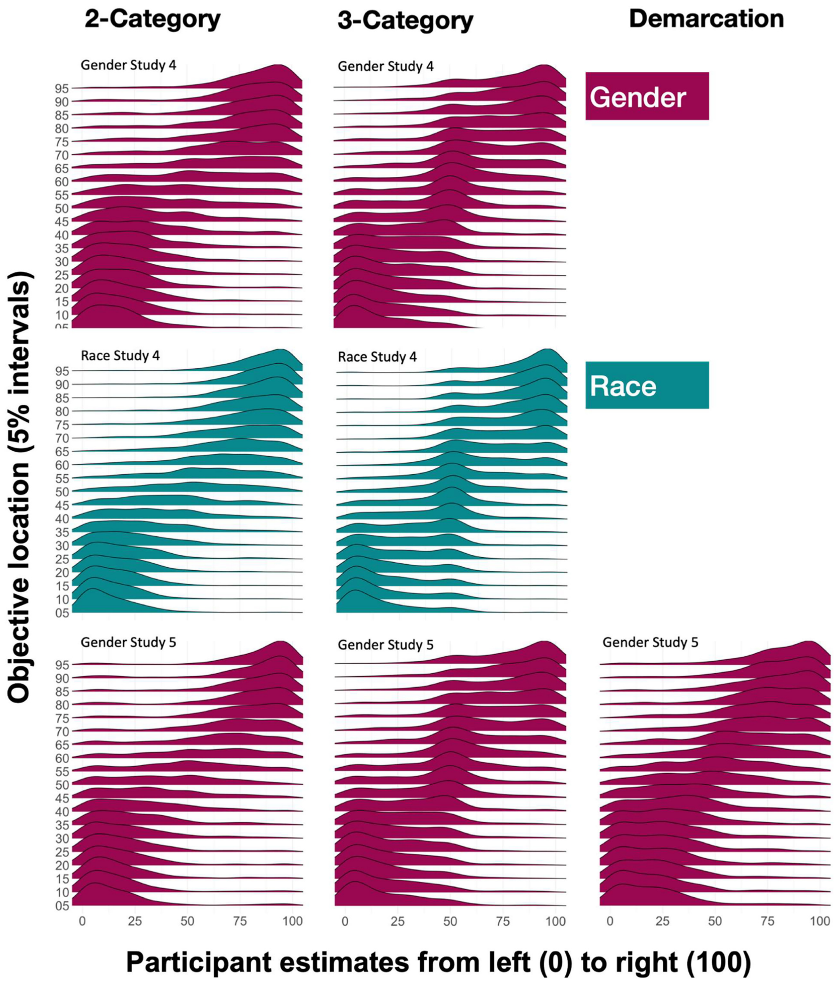
As in previous studies, stimuli were grouped together based on the base face pair used to create them, the location of labels was counterbalanced, task data were collected on Inquisit Web (Millisecond Software, 2016), and we collected the same variables of interest measured in previous studies. Stimuli for Study 5 consisted of the 76 White gender/sex morph faces used as gender/sex stimuli in Study 4.

Results

To determine whether the results differed by condition, we ran a multilevel regression model predicting error from condition, intermediateness, and their interaction along with a random intercept for participant ID and a random slope within-participant for intermediateness. In this analysis, we set the demarcation condition as the reference group. In addition to replicating our Study 4 result of overall less error in the two-category condition than three-category condition ($b = -.066$, $t[320] = -4.33$, $p < .001$), we also found that the demarcation condition differed from both of these conditions. Participants in the demarcation condition showed less error than those in the two-category condition ($b = -.032$, $t[320] = -2.07$, $p = .039$), but more error than those in the three-category condition ($b = .035$, $t[320] = 2.25$, $p = .025$). Importantly, for the most intermediary stimuli (at exactly halfway between the endpoints), participants in the demarcation condition demonstrated more error than those in the three-category condition, $t(825) = 4.00$, $p < .001$, 99% confidence interval (CI) = [0.00617, 0.07515], but no different than

Figure 8

Distribution of Participant Estimates of Stimulus Location Along the Continuum by Objective (Correct) Location in Studies 4 and 5



Note. See the online article for the color version of this figure.

those in the two-category condition, $t(858) = -2.00$, $p > .05$, 99% CI = [−0.0517, 0.0164].

Discussion

These results suggest that providing a third category provides a greater advantage over and above a reference point for participants when categorizing intermediary stimuli. Combined with the results of Study 4, it appears that for cisgender people, adding a new gender/sex label makes people more accurate at judging androgynous faces.

General Discussion

Our results both reinforce and challenge the existing literature on how people sort and categorize faces according to their gender/sex and race. In line with past research (Freeman, Rule, et al., 2010), we found that the relation between the objective location of target faces and participants' placement of those faces along a continuum was nonlinear. In addition, both our work and past work (Campanella et al., 2001) suggest that more prototypical stimuli are seen as fitting into categories; morphed faces that were 70%–90% one gender/sex or race were generally judged similarly to one another and seem to have been sorted categorically.

While supporting the view that categories play a large role in shaping face perception, our results clarify the effects of such categories on faces that do not fall neatly within a single canonical category. In line with previous work, participants' average response to 50% of morphs was accurate (Chen & Hamilton, 2012; Freeman, Rule, et al., 2010). However, it was not clear from past work whether this was due to participants responding bimodally (averaging out to an accurate estimate) or if they conceptualized these intermediate members as they were—in the middle. In contrast to both of these possibilities, we found that (cisgender) participants' responses to intermediary faces were nearly evenly spread across the entire continuum and that participants demonstrated greater error when making judgments about faces that fell between the two canonical categories, an inaccuracy that was previously obscured by averaging responses.

This was particularly true for gender/sex, although data in the domain of race showed a similar (albeit slightly weaker) pattern. Although we did not investigate the rationale behind the difference across domains, it is possible that this is due to differences in conceptualization across the two domains. While gender is overwhelmingly discussed as binary/bimodal (Martin & Slepian, 2021), race is a domain with many subcategories, thus potentially weakening the strength of our bimodal task. Additionally, it could be that our participants have had less exposure to androgynous individuals than bi and/or multiracial individuals, resulting in a more binary/bimodal representation of gender/sex than race.

Finally, our results challenge assumptions about the rigidity of category structure within dimensions of gender/sex and race and build upon past work in the domain of race demonstrating that the presence of a "third" category engenders a shift in participants' explicit categorization of multi- and biracial faces (Chen & Hamilton, 2012; Chen et al., 2023; Gaither et al., 2019; Levy et al., 2023; Peery & Bodenhausen, 2009; Young et al., 2021). Our work extends these findings to a different domain (gender/sex) and demonstrates that this shift occurs even when the intermediary category is labeled with a recognizable term and the response options are continuous. While recent work examining the influence of a third racial category has framed their investigations in terms of possible mechanisms underlying people's racial categorization process (e.g., hypodescent; see Young et al., 2021, essentialism; see Gaither et al., 2019; Levy et al., 2023, or their own racial identity; see Chen et al., 2023), we focus primarily on the accuracy of intermediary stimuli across contexts. Specifically, we find that providing participants with a "third option" results in less error in judgments of the location of intermediary faces than a midpoint mark indicating the center of the response space. Further, while we did measure several possible covariates or predictors of our participants' judgments (e.g., beliefs about gender/sex diversity, transphobia, warmth towards cisgender and transgender people (in cisgender participants), and gender rumination and theoretical awareness of gender (in gender-diverse participants), none yielded significant systematic associations between people's accuracy and their belief about gender (see the online supplemental materials, pp. 15–22). Thus, unlike recent work looking at racial categorization, our data are unable to shed light on the underlying constructs that influence people's performance on our task. This is an important area of continued research moving forward.

Despite a clear advantage of a third category when sorting intermediary stimuli along the dimensions of gender/sex and race, adding a third category resulted in an overall cost in terms of both speed and overall accuracy (see the online supplemental materials, pp. 26–27). In line with past work finding that cognitive load has a negative impact on the accuracy of race perception (Chen & Hamilton, 2012), our interpretation is that this may have been driven by the cost of learning, remembering, and mentally rehearsing three symbol-face pairings (vs. two; Paas & Van Merriënboer, 1994; Sweller, 1988). However, unlike Chen and Hamilton who found that the cognitive load primarily impacted the categorization of intermediate faces, we find that the additional cognitive load resulted in deficits when categorizing all kinds of faces.

The results of our gender-diverse sample further suggest that judging gender/sex as binary and categorical is not inevitable, nor is uncertainty about the location of more androgynous faces. The gender-diverse sample was more accurate than our cisgender samples in their placement of androgynous faces. Combined with our experimental results, we believe one reason for this result may be that gender-diverse people, on average, had more continuous representations of gender/sex and/or trimodal representations of gender/sex. Another possibility is that gender-diverse participants' experience seeing androgynous faces across their lifetime gave them an advantage at categorizing androgynous faces (including, e.g., people in the midst of medical transition) relative to cisgender heterosexual participants. Finally, these participants may have been more motivated to correctly locate androgynous faces along the continuum. In any of these cases, these results further indicate the role that experience plays in our judgments of social dimensions.

Constraints on Generality

In this set of studies, we consistently found that participants demonstrated greater error when judging intermediary stimuli along the dimensions of gender/sex and race than when judging stimuli that were closer to the more prototypical endpoints. Based on the diversity of our stimuli and participants across studies, we believe this effect persists when categorizing morphed faces of different races, gender/sexes, and from several different face databases. We found these results across individuals recruited from different platforms (undergraduate subject pool vs. online) with different racial and gender identities. However, there are several dimensions that we did not examine and therefore our work may not generalize to. First, the participants in these studies were disproportionately American and relatively young adults, thus potentially limiting the generalizability of our findings by nationality and age. If anything, however, we would predict that older adults might have even more rigid ideas about gender/sex, as there's some evidence that younger adults may be thinking about gender/sex differently from previous generations (Wong, 2015). In addition, because racial categories differ depending on one's geographical location (Pauker, Carpinella, et al., 2018), our primarily American sample limits the conclusions we can draw about racial judgments to individuals who are already familiar with common racial categories used in the United States. In addition, we created face stimuli by morphing faces in varying proportions. This allowed for greater control over image characteristics (e.g., camera angle and background color), and had the advantage of enabling us to compare participants' responses to an objective target value. However, it is unknown if our effects would hold when using more ecologically valid stimuli that vary naturally along the dimensions of gender/sex and race (see Crookes et al., 2015; Hehman et al., 2017). Still, our concern that the morphed nature of the stimuli drove the results is lessened by the fact that our gender-diverse sample had less difficulty sorting these same faces and because our experimental manipulation in Studies 4 and 5 worked effectively. Finally, there are many nonfacial features such as clothing, voice, and gait that are often associated with social categories but were not examined in the present study (Cao et al., 2008; Johnson & Tassinary, 2005; Ko et al., 2006). Thus, it is unclear whether or not these findings would hold outside our experimental setting, and future research should continue to unpack how these factors might influence people's judgments of more intermediary or less prototypical members of gender/sex and race groups.

An additional potential constraint to our work is the expansion of our use of terms like "error" or accuracy. In our task, we were comfortable using these terms because the morphed faces were mathematically created to produce "correct" responses in terms of where they fit between two anchors. However, outside of carefully controlled lab settings with novel stimuli, these words have different implications and, in fact, often do not apply to everyday social perception. For instance, it is not possible to take a random person's face and precisely indicate what percentage masculine or feminine it is "objectively" because there is no right answer. Therefore, in studies focused on real human faces, one can only comment about the degree to which a given rater does or does not agree with other raters. Alternatively, accuracy or error can be used in other settings to refer to, for example, the alignment between how a person categorizes a stimulus and how the person being rated categorizes themself. For example, does an evaluator think a person who identifies as a Black–White biracial is a Black–White biracial or in another group? This is an interesting set of questions about identity, but one that our study did not address.

Future Directions

Future research should also continue to consider other types of response options beyond either the discrete categories used in much of the previous research on gender/sex and race perception and the single continuous scale used in the current study. Neither of these may fully capture the nuance with which people can or do judge social categories like gender and race. For instance, there is evidence that people's judgments of masculinity and femininity are better captured by multiple independent scales (Hester et al., 2021). Likewise, Nicolas et al. (2019) found that using a free-response measure led participants to label Black–White mixed-race faces using a greater number of alternative race categories than are typically included in categorization tasks with discrete choices. Another important future direction is to investigate where in the perceptual process changes occur that affect people's judgments of the sex/gender and race of faces. Do either the gender-diverse participants in Study 3 or the participants provided a third label in Studies 4 and 5 actually see faces differently or do differences only emerge when participants are asked to make judgments about faces?

In sum, our results suggest that while cisgender American adults may default to judge faces categorically, how they think about and use social categories can change through experience. This work raises the possibility that both personal experiences and the adoption of new labels may help us appreciate diversity, especially for those who do not readily fit into prototypical social categories.

References

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods & Research*, *25*(3), 318–340. https://doi.org/10.1177/0049124197025003003

Bastian, B., & Haslam, N. (2008). Immigration from the perspective of hosts and immigrants: Roles of psychological essentialism and social identity. *Asian Journal of Social Psychology*, *11*(2), 127–140. https://doi.org/10.1111/j.1467-839X.2008.00250.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv. https://doi.org/10.48550/ARXIV.1406.5823

Bauerband, L. A., & Galupo, M. P. (2014). The gender identity reflection and rumination scale: Development and psychometric evaluation. *Journal of Counseling & Development*, *92*(2), 219–231. https://doi.org/10.1002/j.1556-6676.2014.00151.x

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

Blascovich, J., Wyer, N. A., Swart, L. A., & Kibler, J. L. (1997). Racism and racial categorization. *Journal of Personality and Social Psychology*, *72*(6), 1364–1372. https://doi.org/10.1037/0022-3514.72.6.1364

Bradford, N. J., Rider, G. N., Catalpa, J. M., Morrow, Q. J., Berg, D. R., Spencer, K. G., & McGuire, J. K. (2019). Creating gender: A thematic analysis of genderqueer narratives. *International Journal of Transgenderism*, *20*(2–3), 155–168. https://doi.org/10.1080/15532739.2018.1474516

Brigham, J. C., & Barkowitz, P. (1978). Do "They all look alike?" The effect of race, sex, experience, and attitudes on the ability to recognize faces. *Journal of Applied Social Psychology*, *8*(4), 306–318. https://doi.org/10.1111/j.1559-1816.1978.tb00786.x

Brigham, J. C., & Malpass, R. S. (1985). The role of experience and contact in the recognition of faces of own- and other-race persons. *Journal of Social Issues*, *41*(3), 139–155. https://doi.org/10.1111/j.1540-4560.1985.tb01133.x

Bussey, K. (2011). Gender identity development. In S. J. Schwartz, K. Luyckx, & V. L. Vignoles (Eds.), *Handbook of identity theory and research* (pp. 603–628). Springer. https://doi.org/10.1007/978-1-4419-7988-9_25

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

Cao, L., Dikmen, M., Fu, Y., & Huang, T. S. (2008, October 26). *Gender recognition from body* [Conference session]. In Proceedings of the 16th ACM International Conference on Multimedia, Vancouver British Columbia Canada (pp. 725–728). https://doi.org/10.1145/1459359.1459470

Carroo, A. W. (1986). Other race recognition: A comparison of Black American and African subjects. *Perceptual and Motor Skills*, *62*(1), 135–138. https://doi.org/10.2466/pms.1986.62.1.135

Chen, J. M., & Hamilton, D. L. (2012). Natural ambiguities: Racial categorization of multiracial individuals. *Journal of Experimental Social Psychology*, *48*(1), 152–164. https://doi.org/10.1016/j.jesp.2011.10.005

Chen, J. M., Meyers, C., Pauker, K., Gaither, S. E., Hamilton, D. L., & Sherman, J. W. (2023). Intergroup context moderates the impact of white Americans' identification on racial categorization of ambiguous faces. *Personality and Social Psychology Bulletin*. Advance online publication. https://doi.org/10.1177/01461672231190264

Chen, J. M., Moons, W. G., Gaither, S. E., Hamilton, D. L., & Sherman, J. W. (2014). Motivation to control prejudice predicts categorization of multiracials. *Personality and Social Psychology Bulletin*, *40*(5), 590–603. https://doi.org/10.1177/0146167213520457

Correll, J., Ma, D. S., & Davis, J. P. (2021). Perceptual tuning through contact? Contact interacts with perceptual (not memory-based) face-processing ability to predict cross-race recognition. *Journal of Experimental Social Psychology*, *92*, Article 104058. https://doi.org/10.1016/j.jesp.2020.104058

Crisp, R. J., & Hewstone, M. (2007). Multiple social categorization. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 39, pp. 163–254). Elsevier. https://doi.org/10.1016/S0065-2601(06)39004-1

Crookes, K., Ewing, L., Gildenhuys, J., Kloth, N., Hayward, W. G., Oxner, M., Pond, S., & Rhodes, G. (2015). How well do computer-generated faces tap face expertise? *PLoS One*, *10*(11), Article e0141353. https://doi.org/10.1371/journal.pone.0141353

Davis, G. (2015). *Contesting intersex: The dubious diagnosis*. New York University Press.

DeBruine, L. (2018). *debruine/webmorph: Beta release 2* (Version 0.0.0.9001) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.1162670

DeBruine, L., & Jones, B. (2021). *Face research lab London set (p. 281699214 Bytes)* [Data set]. figshare. https://doi.org/10.6084/M9.FIGSHARE.5047666.V5 de Heering, A., & Rossion, B. (2008). Prolonged visual experience in adulthood modulates holistic face perception. *PLoS One*, *3*(5), Article e2317. https://doi.org/10.1371/journal.pone.0002317

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Dunham, Y., & Olson, K. R. (2016). Beyond discrete categories: Studying multiracial, intersex, and transgender children will strengthen basic developmental science. *Journal of Cognition and Development*, *17*(4), 642–665. https://doi.org/10.1080/15248372.2016.1195388

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*(3), 316–326. https://doi.org/10.1177/0146167297233009

Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist*, *54*(6), 408–423. https://doi.org/10.1037/0003-066X.54.6.408

Factor, R., & Rothblum, E. (2008). Exploring gender identity and community among three groups of transgender individuals in the United States: MTFs, FTMs, and genderqueers. *Health Sociology Review*, *17*(3), 235–253. https://doi.org/10.5172/hesr.451.17.3.235

Fallshore, M., & Schooler, J. W. (1995). Verbal vulnerability of perceptual expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(6), 1608–1623. https://doi.org/10.1037/0278-7393.21.6.1608

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002

Feng, L., Liu, J., Wang, Z., Li, J., Li, L., Ge, L., Tian, J., & Lee, K. (2011). The other face of the other-race effect: An fMRI investigation of the other-race face categorization advantage. *Neuropsychologia*, *49*(13), 3739–3749. https://doi.org/10.1016/j.neuropsychologia.2011.09.031

Firestone, C., & Scholl, B. J. (2015). Can you experience "top-down" effects on perception?: The case of race categories and perceived lightness. *Psychonomic Bulletin & Review*, *22*(3), 694–700. https://doi.org/10.3758/s13423-014-0711-5

Fishbein, H. D., & Imai, S. (1993). Preschoolers select playmates on the basis of gender and race. *Journal of Applied Developmental Psychology*, *14*(3), 303–316. https://doi.org/10.1016/0193-3973(93)90012-K

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). Elsevier. https://doi.org/10.1016/S0065-2601(08)60317-2

Freeman, J. B., Pauker, K., Apfelbaum, E. P., & Ambady, N. (2010). Continuous dynamics in the real-time perception of race. *Journal of Experimental Social Psychology*, *46*(1), 179–185. https://doi.org/10.1016/j.jesp.2009.10.002

Freeman, J. B., Rule, N. O., Adams, R. B., & Ambady, N. (2010). The neural basis of categorical face perception: Graded representations of face gender in fusiform and orbitofrontal cortices. *Cerebral Cortex*, *20*(6), 1314–1322. https://doi.org/10.1093/cercor/bhp195

Gagné, P., Tewksbury, R., & McGaughey, D. (1997). Coming out and crossing over: Identity formation and proclamation in a transgender community. *Gender & Society*, *11*(4), 478–508. https://doi.org/10.1177/089124397011004006

Gaither, S. E. (2015). "Mixed" results: Multiracial research and identity explorations. *Current Directions in Psychological Science*, *24*(2), 114–119. https://doi.org/10.1177/0963721414558115

Gaither, S. E., Chen, J. M., Pauker, K., & Sommers, S. R. (2019). At face value: Psychological outcomes differ for real vs. computer-generated multiracial faces. *The Journal of Social Psychology*, *159*(5), 592–610. https://doi.org/10.1080/00224545.2018.1538929

Gaither, S. E., Pauker, K., & Johnson, S. P. (2012). Biracial and monoracial infant own-race face perception: An eye tracking study: Biracial and monoracial infant face perception. *Developmental Science*, *15*(6), 775–782. https://doi.org/10.1111/j.1467-7687.2012.01170.x

Galupo, M. P., Pulice-Farrow, L., & Pehl, E. (2021). "There is nothing to do about it": Nonbinary individuals' experience of gender dysphoria. *Transgender Health*, *6*(2), 101–110. https://doi.org/10.1089/trgh.2020.0041

Golby, A. J., Gabrieli, J. D. E., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience*, *4*(8), 845–850. https://doi.org/10.1038/90565

Goldstone, R. L., Steyvers, M., & Larimer, K. (1996). Categorical perception of novel dimensions. In G. W. Cottrell (Ed.), *Proceedings of the eighteenth annual conference of the cognitive science society* (p. 6). Routledge.

Harnad, S. (1987). *Psychophysical and cognitive aspects of categorical perception: A critical overview* (S. Harnad, ed.; pp. 1–52). Cambridge University Press. https://cogprints.org/1571/

Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology*, *113*(4), 513–529. https://doi.org/10.1037/pspa0000090

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition*, *21*(9–10), 1306–1336. https://doi.org/10.1080/13506285.2013.823140

Hester, N., Jones, B. C., & Hehman, E. (2021). Perceived femininity and masculinity contribute independently to facial impressions. *Journal of Experimental Psychology: General*, *150*(6), 1147–1164. https://doi.org/10.1037/xge0000989

Højsgaard, S., Halekoh, U., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, *15*(2), 1–11. https://doi.org/10.18637/jss.v015.i02

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: An integrative account of the other-race recognition deficit. *Psychological Review*, *117*(4), 1168–1187. https://doi.org/10.1037/a0020463

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*, *74*(2), 171–193. https://doi.org/10.1037/amp0000307

Iankilevitch, M., Cary, L. A., Remedios, J. D., & Chasteen, A. L. (2020). How do multiracial and monoracial people categorize multiracial faces? *Social Psychological and Personality Science*, *11*(5), 688–696. https://doi.org/10.1177/1948550619884563

Iantaffi, A., & Bockting, W. O. (2011). Views from both sides of the bridge? Gender, sexual legitimacy and transgender people's experiences of relationships. *Culture, Health & Sexuality*, *13*(3), 355–370. https://doi.org/10.1080/13691058.2010.537770

IJzerman, H., & Semin, G. R. (2009). The thermometer of social relations: Mapping social proximity on temperature. *Psychological Science*, *20*(10), 1214–1220. https://doi.org/10.1111/j.1467-9280.2009.02434.x

Ito, T. A., & Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology*, *85*(4), 616–626. https://doi.org/10.1037/0022-3514.85.4.616

Johnson, K. L., & Tassinary, L. G. (2005). Perceiving sex directly and indirectly: Meaning in motion and morphology. *Psychological Science*, *16*(11), 890–897. https://doi.org/10.1111/j.1467-9280.2005.01633.x

Kang, S. K., & Bodenhausen, G. V. (2015). Multiple identities in social perception and interaction: Challenges and opportunities. *Annual Review of Psychology*, *66*(1), 547–574. https://doi.org/10.1146/annurev-psych-010814-015025

Ko, S. J., Judd, C. M., & Blair, I. V. (2006). What the voice reveals: Within- and between-category stereotyping on the basis of voice. *Personality and Social Psychology Bulletin*, *32*(6), 806–819. https://doi.org/10.1177/0146167206286627

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lee, K., Anzures, G., Quinn, P. C., Pascalis, O., & Slater, A. (2011). *Development of face processing expertise*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199559053.013.0039

Levin, D. T. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1364–1382. https://doi.org/10.1037/0278-7393.22.6.1364

Levin, D. T., & Angelone, B. L. (2002). Categorical perception of race. *Perception*, *31*(5), 567–578. https://doi.org/10.1068/p3315

Levin, D. T., & Beale, J. M. (2000). Categorical perception occurs in newly learned faces, other-race faces, and inverted faces. *Perception & Psychophysics*, *62*(2), 386–401. https://doi.org/10.3758/BF03205558

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

Levy, A., Nguyen, C., Slepian, M. L., Gaither, S., Pauker, K., & Dovidio, J. F. (2023). Categorizing a face and facing a category: The constructive impacts of ambiguity and uncertainty in racial categorization. *Personality and Social Psychology Bulletin*, *49*(6), 910–924. https://doi.org/10.1177/01461672221084537

Lindsay, D. S., Jack, P. C., & Christian, M. A. (1991). Other-race face perception. *Journal of Applied Psychology*, *76*(4), 587–589. https://doi.org/10.1037/0021-9010.76.4.587

LoBue, V., & Thrasher, C. (2015). The child affective facial expression (CAFE) set: Validity and reliability from untrained adults. *Frontiers in Psychology*, *5*, Article 1532. https://doi.org/10.3389/fpsyg.2014.01532

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

Ma, Y., Yang, S., & Han, S. (2011). Attitudes influence implicit racial face categorization in a perceptual task. *Group Processes & Intergroup Relations*, *14*(6), 887–899. https://doi.org/10.1177/1368430211409962

MacLin, O. H., & Malpass, R. S. (2001). Racial categorization of faces: The ambiguous race face effect. *Psychology, Public Policy, and Law*, *7*(1), 98–118. https://doi.org/10.1037/1076-8971.7.1.98

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, *13*(4), 330–334. https://doi.org/10.1037/h0028434

Martin, A. E., & Slepian, M. L. (2021). The primacy of gender: Gendered cognition underlies the big two dimensions of social cognition. *Perspectives on Psychological Science*, *16*(6), 1143–1158. https://doi.org/10.1177/1745691620904961

McGuire, J. K., Beek, T. F., Catalpa, J. M., & Steensma, T. D. (2019). The Genderqueer Identity (GQI) Scale: Measurement and validation of four distinct subscales with trans and LGBQ clinical and community samples in two countries. *International Journal of Transgenderism*, *20*(2–3), 289–304. https://doi.org/10.1080/15532739.2018.1460735

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037/1076-8971.7.1.3

Meissner, C. A., Brigham, J. C., & Butz, D. A. (2005). Memory for own- and other-race faces: A dual-process approach. *Applied Cognitive Psychology*, *19*(5), 545–567. https://doi.org/10.1002/acp.1097

Millisecond Software. (2016). *Inquisit 5* [Computer software].

Nagoshi, J. L., Brzuzy, S., & Terrell, H. K. (2012). Deconstructing the complex perceptions of gender roles, gender identity, and sexual orientation among transgender individuals. *Feminism & Psychology*, *22*(4), 405–422. https://doi.org/10.1177/0959353512461929

Nagoshi, J. L., & Nagoshi, C. T. (2008). *Transphobia scale* [Data set]. American Psychological Association. https://doi.org/10.1037/t12384-000

Nicolas, G., Skinner, A. L., & Dickter, C. L. (2019). Other than the sum: Hispanic and middle eastern categorizations of Black–White mixed-race faces. *Social Psychological and Personality Science*, *10*(4), 532–541. https://doi.org/10.1177/1948550618769591

Olian, J. D., Schwab, D. P., & Haberfeld, Y. (1988). The impact of applicant gender compared to qualifications on hiring recommendations. *Organizational Behavior and Human Decision Processes*, *41*(2), 180–195. https://doi.org/10.1016/0749-5978(88)90025-8

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*(4), 351–371. https://doi.org/10.1007/BF02213420

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology*, *34*(1), 181–209. https://doi.org/10.1146/annurev.soc.33.040406.131740

Pauker, K., & Ambady, N. (2009). Multiracial faces: How categorization affects memory at the boundaries of race. *Journal of Social Issues*, *65*(1), 69–86. https://doi.org/10.1111/j.1540-4560.2008.01588.x

Pauker, K., Carpinella, C. M., Lick, D. J., Sanchez, D. T., & Johnson, K. L. (2018). Malleability in biracial categorizations: The impact of geographic context and targets' racial heritage. *Social Cognition*, *36*(5), 461–480. https://doi.org/10.1521/soco.2018.36.5.461

Pauker, K., Meyers, C., Sanchez, D. T., Gaither, S. E., & Young, D. M. (2018). A review of multiracial malleability: Identity, categorization, and shifting racial attitudes. *Social and Personality Psychology Compass*, *12*(6), Article e12392. https://doi.org/10.1111/spc3.12392

Pauker, K., Rule, N. O., & Ambady, N. (2010). Ambiguity and social perception. In E. Balcetis & G. D. Lassiter (Eds.), *Social psychology of visual perception* (pp. 7–26). Psychology Press.

Pauker, K., Weisbuch, M., Ambady, N., Sommers, S. R., Adams, R. B., & Ivcevic, Z. (2009). Not so black and white: Memory for ambiguous group members. *Journal of Personality and Social Psychology*, *96*(4), 795–810. https://doi.org/10.1037/a0013265

Peery, D., & Bodenhausen, G. V. (2009). Ambiguity and ambivalence in the voting booth and beyond: A social-psychological perspective on racial attitudes and behavior in the Obama era. *Du Bois Review: Social Science Research on Race*, *6*(1), 71–82. https://doi.org/10.1017/S1742058X09090067

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741–763. https://doi.org/10.1037/0022-3514.67.4.741

Quinn, P. C., Yahr, J., Kuhn, A., Slater, A. M., & Pascalis, O. (2002). Representation of the gender of human faces by infants: A preference for female. *Perception*, *31*(9), 1109–1121. https://doi.org/10.1068/p3331

Reis, E. (2021). *Bodies in doubt: An American history of intersex* (2nd ed.). Johns Hopkins University Press.

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

Ridgeway, C. L. (2011). *Framed by gender: How gender inequality persists in the modern world*. Oxford University Press.

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences*, *50*(1), 90–94. https://doi.org/10.1016/j.paid.2010.09.004

Ruble, D. N., Martin, C. L., Berenbaum, S. A. (2007). Gender development. In N. Eisenberg (Ed.), *Handbook of Child Psychology* (6th ed., Vol. 3, pp. 858–932). https://doi.org/10.1002/9780470147658.chpsy0314

Samal, A., Subramani, V., & Marx, D. (2007). Analysis of sexual dimorphism in human face. *Journal of Visual Communication and Image Representation*, *18*(6), 453–463. https://doi.org/10.1016/j.jvcir.2007.04.010

Schudson, Z. C., & van Anders, S. M. (2022). Gender/sex diversity beliefs: Scale construction, validation, and links to prejudice. *Group Processes & Intergroup Relations*, *25*(4), 1011–1036. https://doi.org/10.1177/1368430220987595

Shih, M., & Sanchez, D. T. (2009). When race becomes even more complex: Toward understanding the landscape of multiracial identity and experiences. *Journal of Social Issues*, *65*(1), 1–11. https://doi.org/10.1111/j.1540-4560.2008.01584.x

Simon, D., Chen, J. M., Sherman, J. W., & Calanchini, J. (2023). A recognition advantage for members of higher-status racial groups. *British Journal of Psychology*, *114*(S1), 188–211. https://doi.org/10.1111/bjop.12587

Sugden, N. A., & Marquis, A. R. (2017). Meta-analytic review of the development of face discrimination in infancy: Face race, face gender, infant age, and methodology moderate face discrimination. *Psychological Bulletin*, *143*(11), 1201–1244. https://doi.org/10.1037/bul0000116

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Tanaka, J. W., Kiefer, M., & Bukach, C. M. (2004). A holistic account of the own-race effect in face recognition: Evidence from a cross-cultural study. *Cognition*, *93*(1), B1–B9. https://doi.org/10.1016/j.cognition.2003.09.011

Taylor, M. G., Rhodes, M., & Gelman, S. A. (2009). Boys will be boys; cows will be cows: Children's essentialist reasoning about gender categories and animal species. *Child Development*, *80*(2), 461–481. https://doi.org/10.1111/j.1467-8624.2009.01272.x

van Anders, S. M. (2015). Beyond sexual orientation: Integrating gender/sex and diverse sexualities via sexual configurations theory. *Archives of Sexual Behavior*, *44*(5), 1177–1213. https://doi.org/10.1007/s10508-015-0490-8

Vries, K. M. (2012). Intersectional identities and conceptions of the self: The experience of transgender people. *Symbolic Interaction*, *35*(1), 49–67. https://doi.org/10.1002/symb.2

Walker, P. M., & Tanaka, J. W. (2003). An encoding advantage for own-race versus other-race faces. *Perception*, *32*(9), 1117–1125. https://doi.org/10.1068/p5098

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*(6), 1049–1062. https://doi.org/10.1037/0022-3514.67.6.1049

Weisman, K., Johnson, M. V., & Shutts, K. (2015). Young children's automatic encoding of social categories. *Developmental Science*, *18*(6), 1036–1043. https://doi.org/10.1111/desc.12269

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*(9), 1115–1125. https://doi.org/10.1177/01461672992512005

Wild, H. A., Barrett, S. E., Spence, M. J., O'Toole, A. J., Cheng, Y. D., & Brooke, J. (2000). Recognition and sex categorization of adults' and children's faces: Examining performance in the absence of sex-stereotyped cues. *Journal of Experimental Child Psychology*, *77*(4), 269–291. https://doi.org/10.1006/jecp.1999.2554

Willadsen-Jensen, E. C., & Ito, T. A. (2008). A foot in both worlds: Asian Americans' perceptions of Asian, White, and racially ambiguous faces. *Group Processes & Intergroup Relations*, *11*(2), 182–200. https://doi.org/10.1177/1368430207088037

Williams, M. J., & Eberhardt, J. L. (2008). Biological conceptions of race and the motivation to cross racial boundaries. *Journal of Personality and Social Psychology*, *94*(6), 1033–1047. https://doi.org/10.1037/0022-3514.94.6.1033

Wilson, B. D. M., & Meyer, I. H. (2021). *Nonbinary LGBTQ adults in the United States* [Brief]. UCLA School of Law Williams Institute. https://williamsinstitute.law.ucla.edu/wp-content/uploads/Nonbinary-LGBTQ-Adults-Jun-2021.pdf

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian Blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780–7785. https://doi.org/10.1073/pnas.0701644104

Wong, C. M. (2015). *50 percent of millennials believe gender is a spectrum, fusion's massive millennial poll finds*. Huffpost. https://www.huffpost.com/entry/fusion-millennial-poll-gender_n_6624200

Young, D. M., Sanchez, D. T., Pauker, K., & Gaither, S. E. (2021). A meta-analytic review of hypodescent patterns in categorizing multiracial and racially ambiguous targets. *Personality and Social Psychology Bulletin*, *47*(5), 705–727. https://doi.org/10.1177/0146167220941321

Zax, A., Williams, K., Patalano, A. L., Slusser, E., Cordes, S., & Barth, H. (2019). What do biased estimates tell us about cognitive processing? Spatial judgments as proportion estimation. *Journal of Cognition and Development*, *20*(5), 702–728. https://doi.org/10.1080/15248372.2019.1653297

Zhao, L., & Bentin, S. (2008). Own- and other-race categorization of faces by race, gender, and age. *Psychonomic Bulletin & Review*, *15*(6), 1093–1099. https://doi.org/10.3758/PBR.15.6.1093

Zhou, K., Mo, L., Kay, P., Kwok, V. P. Y., Ip, T. N. M., & Tan, L. H. (2010). Newly trained lexical categories produce lateralized categorical perception of color. *Proceedings of the National Academy of Sciences*, *107*(22), 9974–9978. https://doi.org/10.1073/pnas.1005669107

Received January 31, 2023
Revision received September 17, 2023
Accepted September 30, 2023 ■