**From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

Doctoral Dissertations, 2011-2026 Psychology

2025

# Beyond binary gender: From perceptions to policy decisions

Atwood, S.

Princeton, NJ : Princeton University

http://arks.princeton.edu/ark:/88435/dsp015138jj284

*U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf*

# Beyond binary gender: From perceptions to policy decisions

**S. Atwood**

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY

THE DEPARTMENT OF PSYCHOLOGY

Adviser: Kristina R. Olson

January, 2025

© Copyright by S. Atwood, 2024. All rights reserved.

**Abstract**

As scientists and as a society, we often describe and conduct research under the assumption that gender is binary and categorical (Martin & Slepian, 2021; van Anders, 2015). However, these categories do not account for the growing number of people identifying outside the binary (Rubin et al., 2020). Drawing from social and cognitive psychology, this dissertation investigates how we think about gender outside a binary framework. First, we explore how people perceive individual faces that vary in gendered appearance using a spectrum measure. Across 6 studies, we find that cisgender participants typically sort faces into two binary categories and show less consensus when sorting faces with intermediate gendered appearances. However, this pattern of categorization can be influenced by long-term lived experience (e.g., being transgender or nonbinary) and short-term experimental manipulations (e.g., learning about novel groups). The second paper builds on previous work (Burke et al., 2023) to examine the perception of nonbinary individuals as a group. Across 4 studies, we find that cisgender people view nonbinary individuals as less socially legitimate, less stable, and less likable than other gender groups (e.g., men and women). Additionally, we find that viewing nonbinary individuals as less real, stable, and likable is negatively associated with endorsement of nonbinary-inclusive policies. Finally, we address a specific context where cisgender perceptions of nonbinary people impact access to gender-affirming medical care for those with binary versus nonbinary identities. Here, cisgender people often prioritize binary trans individuals, believing they are more likely to want surgery than nonbinary individuals or those who came out later in life. As a result, cisgender people risk deprioritizing the needs of nonbinary individuals in crucial policy decisions. In sum, this dissertation demonstrates how people conceptualize those outside the binary default and begins to explore how these conceptualizations have practical implications for nonbinary individuals.

Table of Contents

**Abstract** ..... **3**
**Chapter 1: Introduction** ..... **6**
Gendered Face Perception ..... 8
Beliefs about nonbinary gender ..... 10
Beliefs about nonbinary people in the domain of gender affirming healthcare ..... 12
Overview of Dissertation Chapters ..... 14
References ..... 17
**Chapter 2: Flexibility in continuous judgements of gender/sex and race** ..... **31**
Abstract ..... 32
Statement of Public Significance ..... 33
Introduction ..... 34
Studies 1a & 1b ..... 43
Study 2 ..... 56
Study 3 ..... 60
Study 4 ..... 67
Study 5 ..... 74
General Discussion ..... 77
References ..... 85
Supplementary Material ..... 100
**Chapter 3: Perceptions of nonbinary people as an intermediate group** ..... **139**
Abstract ..... 140
Introduction ..... 141
Pilot Study ..... 146
Studies 1a-1c ..... 147
Study 2 ..... 154
General Discussion ..... 162
References ..... 165
Supplementary Material ..... 174
**Chapter 4: Cisgender people's perceptions of candidates for gender affirming surgery** ..... **183**
Abstract ..... 184
Introduction ..... 185
Studies 1a-1c ..... 189
Study 2 ..... 200
Study 3 ..... 205
General Discussion ..... 210
References ..... 215
Supplementary Material ..... 225

**Chapter 5: General Discussion..... 246**
Challenging the inevitability of binary face categorization.....246
Exploring conceptualizations of nonbinary people.....249
Investigating facilitators and limitations of nonbinary affirmation and support..... 251
General Limitations and Future Directions.....254
Conclusion..... 256
References.....258
**Acknowledgements..... 262**

From the moment we are born to the day we die, gender is a core feature of human experience with implications for cognition, social behavior, and development across the lifespan. Historically, both scientists have relied on terms like male/man and female/woman, under the assumption that they sufficiently capture our understanding of gender (Martin & Slepian, 2021). Yet, these categories do not account for the way we conceptualize those who identify outside the gender binary (including but not limited those who identify as nonbinary, genderfluid, genderqueer, agender, and many others see Tate et al., 2014; Matsuno & Budge, 2017; Richards et al., 2017; Monro, 2019; Rubin et al., 2020; Schudson & Morgenroth, 2022). Research suggests that the number of gender diverse individuals has grown substantially in the past decade, with over 11% of LGBTQ+ adults and 26% of LGBTQ+ youth now identifying outside the gender binary (Wilson & Meyer, 2021; The Trevor Project, 2021). This is also reflected at the societal level with increased awareness (Anders, 2021), and (some) accommodation of gender diversity (e.g., recognition of singular they; Harmon, 2019, options for X gender marker on documentation; Blinken, 2022, nonbinary categories in sports, Hernandez, 2023, etc.) is greater than ever before in Western society (Elias, 2019). However, our approach to gender in mainstream social psychology research often fails to acknowledge or account for the growing portion of the population that identifies outside of the binary.

The field of social psychology has historically relied on a binary conception of gender, operating under the assumption that gender is a fixed and universal category that coincides with one's assigned sex at birth (Martin & Slepian, 2021; van Anders, 2015 cf, Hyde et al., 2019). These assumptions manifest in the topics psychologists study such as sex differences (which focuses on the ways males and females are distinct from one another; Maccoby, 1966; Kimura, 1992; Buss, 1995), gender roles (concerning the functions and responsibilities women and men

are expected to play in society; Eagly, 1978; Ridgeway 2001), gender stereotypes (or commonly held beliefs about men and women; Eagly & Steffen, 1984; Swim & Sanna; 1996; Rudman & Glick, 2001), gender norms (focusing on the expectations about how women and men are expected to behave; Bem, 1974; Eagly & Wood 1999; Connell & Messerschmidt, 2005), ambivalent sexism (which encompasses both benevolent and hostile sexism; Glick & Fiske, 2001; Bareket & Fiske, 2023), and gender dynamics in the workplace (e.g. women in leadership; power hierarchies between men and women, etc; Heilman, 2001; Eagly & Carli, 2007; Schmader, 2023). Assumptions about the universality of the gender binary also manifest not only in the topics of research but also the ways research is conducted in the field of psychology. For instance, the most well-known stimuli sets consist of exclusively men's and women's faces (e.g. Chicago Face Database; Ma et al., 2015, London Face Database; DeBruine & Jones, 2021), our most popular measures and scales are designed to refer exclusively to binary women and men, and the vast majority of our study participants identify as binary men and women. On the rare occasion when nonbinary identity is included (most often as the object of study), it's often combined with other categories (e.g., transgender identities) (Burke, Jaurique et al., 2023; Matsuno & Budge, 2017). As a result of conducting research from a gender binarist perspective, ideas about, representations of, and the voices of individuals outside the gender binary are largely absent from mainstream social psychology. Thus, it remains unclear how we perceive, think about, or make decisions about the growing number of individuals who fall outside the binary and also how accounting for gender outside the binary may influence the psychology of gender overall.

This dissertation endeavors to break away from a binarist perspective on gender and addresses how we categorize, think, and make decisions about for individuals and groups outside

the gender binary, with a specific focus on nonbinary people. While there is no single universally agreed upon definition of nonbinary (see Richards et al., 2016; Matsuno & Budge, 2017; Rubin et al., 2020 for representative examples), within this dissertation, nonbinary individuals whose gender identity does not exclusively align with the categories of man or woman. Thus, my co-authors and I investigate how cisgender people conceptualize and respond to individuals and groups whose gendered appearances, identities, and experiences fall between (or beyond) male/man/masculine and female/woman/feminine. First, I extend previous work in face perception and ask how people perceive and categorize faces whose gendered appearance is more intermediate (i.e. less binary). Second, I focus specifically on nonbinary people and ask how cisgender people think about nonbinary people, and explore how people's group level perceptions are related to their policy priorities. Finally, I address one of the potential real life consequences of having a nonbinary gender identity through the lens of cisgender people's decisions about gender affirming medical care. While conducting this research, we sought to make methods more inclusive by using gender-androgynous stimuli (study 1), asking about groups beyond men and women (all studies), and including perspectives of nonbinary participants (study 1).

**Gendered Face Perception**

Past work in face perception has found that people can categorize stereotypically masculine faces as men/male and stereotypically female faces as women/female quickly, accurately, and reliably across people of different races, ethnicities, and ages (Taylor et al., 1978; Devine, 1989; Fiske & Neuberg, 1990; Wild et al., 2000; Weisman et al., 2015). The categorization of faces into two gender categories is thought to develop as early as 3 months old (Quinn et al., 2002), and occurs automatically in adults a mere 150 ms after seeing a face (Ito &

Urland, 2003). This process is so ingrained that people can even categorize faces by gender at above chance levels when faces are inverted (Yang & Dunham, 2019), when face parts are scrambled (Zhao & Hayward, 2010), or when the tops and bottoms halves of faces are misaligned (Chen et al., 2018).

Past research has posited two possible systems through which gender categorization of facial stimuli occurs. First, theories about bottom-up processing are thought to be driven by perceptual processing of specific facial features (e.g. eyes, eyebrows, jaw, and mouth; Roberts & Bruce, 1988; Brown & Perrett, 1993). For example, men have a more square jaw, thicker eyebrows, and a more prominent brow ridge, while women's faces are rounder and more youthful in appearance (Berry & McArthur, 1986). By contrast, theories of top-down processing suggest that gender face categorization is driven by pre-existing social category knowledge (e.g. the beliefs, stereotypes, and/or expectations of the perceiver) as well as situational factors (e.g. interpersonal context and/or the perceivers' motivations; Fiske & Neuberg, 1990; Macrae & Bodenhausen, 2000). For instance, we may rely on our pre-existing knowledge of gender expression (e.g. appearance, mannerisms, roles, etc.) and gender identity (e.g. men and women; Tate et al., 2020) to facilitate gender face categorization. However, these processes are not mutually exclusive. For instance, the Dynamic Interactive Theory of Person Construal has posited that person perception is a continuous process in which both bottom-up and top-down information simultaneously contribute to categorization of a target (Freeman & Ambady, 2011; Kawakami et al., 2017; Freeman et al., 2020).

A plurality of the evidence suggesting that gender categorization is binary is difficult to interpret because the majority of work has relied on frameworks that only allow for binary response options (i.e., participants are asked to label the face as male or female), and/or stimuli

that appear stereotypically male or female in appearance (Roberts & Bruce, 1988; Brown & Perrett, 1993; O'Toole et al., 1998; Kranz & Ishai, 2006; Dobs et al., 2019). The few studies that deviate from this norm have suggested that people's categorizations are still rooted in the gender binary. For instance, Campanella and colleagues (2001), found that when categorizing faces that morphed gradually from a male to female in appearance, participants' judgments abruptly shifted from overwhelmingly male to overwhelmingly female (Campanella et al., 2001). Other work suggests that in addition to perceiving these faces as categorical, people may even exhibit an androcentric bias, and be more likely to judge individuals with ambiguous gender as male (Johnson et al., 2012). Additionally, Freeman and colleagues (2010) found that even when given the option rate faces that varied in gender expression along a 9-point scale from "extremely male" to "extremely female," participants' responses were biased toward the endpoints of the scale, and participants were both slower and less accurate when categorizing the most androgynous faces. While these findings align with research demonstrating that people have worse memory for and are slower at processing non-prototypical individuals (in the domains of race and gender; Fiske et al., 1987, Posner and Keele, 1970), it still leaves open the question of how the most androgynous faces are interpreted.

The first section of the this dissertation extends our knowledge of gender face perception by investigating how intermediate[1] (or androgynous) faces are categorized when presented on a continuous response framework, with the goal of examining how people's categorization patterns for intermediate faces vary on the basis of long-term (lived) experience as a trans or nonbinary person, and short-term experimental manipulations.

**Beliefs about nonbinary gender**

[1] Although we operationalize intermediacy as androgynous appearance for the purposes of this research, we recognize that gendered appearance is not necessarily indicative of gender identity. For this and other reasons, our findings may not reflect the ways people categorize gender diverse individuals in everyday life.

Beliefs about (cisgender) men and women are well-established in the literature and are embedded in the fabric of our everyday lives. A wealth of research has found stereotypes of women as communal, warm, kind, and feminine, and stereotypes of men as agentic, dominant, independent, and masculine (Bem, 1974; Eagly & Steffen, 1984; Swim & Sanna; 1996; Rudman & Glick, 2001; Eagly et al., 2020; Bhatia & Bhatia, 2021; Charlesworth et al., 2021). In contrast to research on men and women, research on the conceptualization of nonbinary people is scant, and what little we know suggests that nonbinary people are perceived as generally unfavorably. For instance, existing work has found that nonbinary people are seen as less attractive, lower status, and less competent than cisgender people (Mao et al., 2019; Jacques et al., 2022). Additionally, other work suggests that we may think of nonbinary people similar to how we think about transgender people. For instance, Perez-Arche & Miller (2021) found that predictors of negative attitudes towards trans people also predict negative attitudes towards nonbinary people. This overlap in perception makes sense, as some nonbinary individuals also identify as transgender (James et al., 2016). However, this perspective may overlook other important dimensions of nonbinary identity as well as a plurality of nonbinary people who do not self-identify as transgender (Wilson & Meyer, 2021). To date, little work has been done to disentangle nonbinary identities  from the larger umbrella category of transgender (Burke et al., 2023), and as a result, little to no (quantitative) research on whether or how nonbinary people are represented as distinct from more binary transgender people (Matsuno & Budge 2017, Schudson & Morgenroth, 2022).

What little we do know about how people think about nonbinary people suggests they may be perceived as intermediate groups (i.e. those that fall between more commonly acknowledged groups on the same dimension of social identity; Burke et al., 2023). Research on

intermediate groups in the domains of sexuality and race has found that people view bisexual and biracial people as less valid or conceptually legitimate in comparison to more established social groups such as gay people, White people, etc (Burke et al., 2023). These intermediate groups are generally liked less than more well-established groups (Herek, 2002; Sanchez & Bonam, 2009; Tate et al., 2014). Finally, some evidence in the domain of sexuality suggests that intermediate groups may be perceived as less stable than more well established groups. For instance, bisexual identity is viewed as an 'unstable' identity adopted as a result of 'confusion' about one's 'true' sexuality (Mohr & Rochlen, 1999; Worthen, 2013; Burke & LaFrance, 2018).

This second section of this dissertation builds on this framework of intermediate groups by first establishing nonbinary as an intermediate group in the domain of gender and then examining the extent to which nonbinary identities are presumed to be less stable, less socially legitimate, and/or that transgender people are less well liked than a range of other gender groups (e.g. cisgender men, cisgender women, transgender men, transgender women, etc.) in the eyes of cisgender heterosexual people. Finally, we address correlates of these conceptualizations by examining if people's ideas of nonbinary people are related to their support for nonbinary affirming policies. In sum, this line of work examines how cisgender, heterosexual people think about nonbinary people and looks at how relate to support for policies impacting nonbinary people's rights and well-being.

**Beliefs about nonbinary people in the domain of gender affirming healthcare**

Beliefs about social groups are impactful because they are often (but not always) aligned with opinions on policies that affect those groups (Pratto et al., 1998; Swim et al., 1995; Henry & Sears, 2002;  cf, see Krysan, 2000; Tuch & Hughes, 2011; Taylor & Parcel, 2019). In the case of gender, previous research has shown that individuals with higher levels of hostile sexism

(reflecting strong negative attitudes towards women) are less likely to support women's rights (Masser & Abrams, 1999; Campbell et al., 1997). Similarly, those who support transgender rights often cite their positive attitudes towards trans people as a key factor in their advocacy for trans-inclusive policies (Morgenroth, Axt, & Westgate, 2024), while individuals with higher levels of transphobia tend to oppose transgender rights (Miller, et al., 2020; Flores et al., 2018). In the final section of this dissertation, we examine cisgender heterosexual people's beliefs about access to gender affirming healthcare for trans and nonbinary people.

Past research suggests that trans and nonbinary people face unique barriers within the healthcare system that make it difficult for them to access gender-affirming care (Houssayni & Nilsen, 2018; dickey et al., 2016; Bakko & Kattari, 2020; Kattari & Hasche, 2016; Lefevor et al., 2019; Kattari et al., 2019; Safer et al., 2016; Lerner & Robles, 2017). Some of these barriers include a lack of gender-affirming care providers (Puckett et al., 2018; Cohen et al., 2020; Sevelius & Jenness, 2017; Clark et al., 2018), meeting eligibility criteria to have surgery (Coleman et al., 2022; Cavanaugh et al., 2016; Ashley, 2019), and difficulty obtaining insurance approval (Downing et al., 2022; Chu et al., 2024). As a result, surgeons, insurance adjusters, and mental health professionals - individuals who tend to be cisgender - play a unique role in deciding who has access to gender affirming care.

Given the impact of cisgender people's decisions in allocating healthcare to trans and nonbinary individuals, it is crucial to understand what goes into making these decisions. Specifically, we aim to uncover the factors cisgender people consider when making decisions about who should access gender affirming healthcare. In particular we focus on the extent to which cisgender people prioritize characteristics most commonly associated with binary transgender individuals (e.g. those who come out earlier, have a gender expression that aligns

with a binary gender, and who self-identify with a binary gender label) over characteristics associated with nonbinary individuals (e.g. those who come out relatively later in life, retain aspects of gender expression associated with their birth-assigned gender, and self-identify as being nonbinary). In addition, we seek to understand the rationale behind why cisgender people choose to prioritize individuals with some characteristics over others, and explore the possibility that cisgender people prioritize surgery candidates who they believe have a higher desire for surgery. In sum, this line of research looks at how cisgender people make decisions about who to prioritize for gender affirming care and considers how these decisions may influence differently effect binary trans vs. nonbinary individuals who seek gender affirming surgery.

**Overview of Dissertation Chapters**

In this dissertation, I build off previous research in face perception, social categorization, and existing beliefs about gender groups to explore how we conceptualize (sections 1 & 2) and make decisions (section 3) about individuals (sections 1 & 3) and groups (section 2) whose gender falls outside of the binary.

First, I build off previous work in social categorization and ask how people perceive and categorize faces whose gendered appearance is more intermediate (i.e. less binary). Using a continuous measure that allowed participants to place stimuli along a continuum from male to female, my collaborators and I found that cisgender participants default to sorting binary gendered faces in a bimodal fashion and showed less consensus when placing faces of more intermediate gendered appearances. However, by expanding our participant group to include gender minorities (i.e. individuals who self-identified transgender and nonbinary) demonstrated people of other gender groups don't default to the same sorting patterns as cisgender people. Finally, we demonstrated that this categorization pattern can also be changed by the short-term

experience of learning about a third category in a perceptual task. In sum, we find that people's perceptions of intermediate gender are malleable and may be influenced by both well-known labels as well as personal experience.

In section 2, we use a series of online studies to explore how cisgender and heterosexual people think about nonbinary people as a group. Building off of previous literature on people with 'intermediate' groups (i.e. groups who are characterized as being 'in between' other groups along the same identity dimension; Burke et al., 2023), we first find that nonbinary is thought of as being an intermediate group. We then explore whether nonbinary people are (like other intermediate groups) thought of as less socially legitimate, that their identities are perceived as less stable, and that they are less likable than more well-known gender groups (e.g. men and women). Finally, we investigate whether or not the cisgender hetersexual's views of nonbinary people are associated with their endorsement of inclusive and affirming policies that would (if approved) tangibly improve nonbinary people's lives. Collectively, our findings demonstrate that nonbinary individuals may not be recognized as a legitimate group in society, and that this may hinder people's support for policies that affirm nonbinary people.

Finally, in section 3, we examine the potential impact of cisgender people's perceptions on the lives of nonbinary people in the domain of gender-affirming healthcare. Using a series of vignette studies, we ask how cisgender people prioritize surgery seekers for gender affirming care when demand (e.g. desire for gender affirming surgery) outpaces supply (e.g. the number of surgeons who are qualified to perform gender affirming surgery). We find that when asked who should get gender affirming surgery, cisgender people often prioritize individuals whose identity development and gender expression aligns with common representations of people who access surgery (e.g. binary trans people, people who came out earlier in life, and those who have a more

binary presentation). Probing further as to why this might be the case, we find that participants rated binary trans individuals and people who had come out at earlier ages as more likely to want surgery than nonbinary individuals and those who came out at later ages. Nonetheless, when desire for surgery was described as equivalent, cisgender, heterosexual participants still deprioritized nonbinary people in recommendations for surgery. Collectively, these findings suggest people may use their intuitions about who wants gender affirming care to decide who should be prioritized for it, and as a result deprioritize people with nonbinary gender identities.

This dissertation explores how we think about nonbinary gender and examines how these understandings influence attitudes and decisions about policies that have practical implications for nonbinary individuals. Looking across studies, our findings suggest that binary conceptions of gender underlie how cisgender individuals think about all people's gender, including those who identify outside of the gender binary. As a result, nonbinary people may be thought of as less well-defined than other gender groups in the eyes of cisgender people. On one hand, being perceived as less well-defined may contribute to cisgender people's biases about nonbinary people, such as perceiving them as a less legitimate group. This, in turn, is associated with negative outcomes for nonbinary people, such as the deprioritization of nonbinary-affirming policies. On the other hand, being a less well-defined group may also have some advantages, such as increased malleability in how they are perceived. For instance, alternative frameworks or ideas about gender can shift how we think about gender outside the binary, even in the short term. Ultimately, understanding how we think about and conceptualize gender-diverse identity and the circumstances under which we prioritize the well-being of gender-diverse individuals will inform both our evolving understanding of gender as well as our approach to discourse around gender-affirming social policies.

## References


Anders, C. (2021, June). More than 1 million nonbinary adults live in the U.S., a pioneering study finds. *The Washington Post*. https://www.washingtonpost.com/dc-md-va/2021/06/22/first-population-estimate-lgbtq-non-binary-adults-us-is-out-heres-why-that-matters/

Ashley, F. (2019). Gatekeeping hormone replacement therapy for transgender patients is dehumanising. *Journal of Medical Ethics*, *45*(7), 480–482. https://doi.org/10.1136/medethics-2018-105293

Bakko, M., & Kattari, S. K. (2020). Transgender-Related Insurance Denials as Barriers to Transgender Healthcare: Differences in Experience by Insurance Type. *Journal of General Internal Medicine*, *35*(6), 1693–1700. https://doi.org/10.1007/s11606-020-05724-2

Bareket, O., & Fiske, S. T. (2023). A systematic review of the ambivalent sexism literature: Hostile sexism protects men's power; benevolent sexism guards traditional gender roles. *Psychological Bulletin*. https://doi.org/10.1037/bul0000400

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42*(2), 155–162. https://doi.org/10.1037/h0036215

Berry, D. S., & McArthur, L. Z. (1986). Perceiving character in faces: The impact of age-related craniofacial changes on social perception. *Psychological Bulletin*, *100*(1), 3–18. https://doi.org/10.1037/0033-2909.100.1.3

Bhatia, N., & Bhatia, S. (2021). Changes in Gender Stereotypes Over Time: A Computational Analysis. *Psychology of Women Quarterly*, *45*(1), 106–125. https://doi.org/10.1177/0361684320977178

Blinken, A. J. (2022, March). *X Gender Marker Available on U.S. Passports Starting April 11* [Online post]. US Department of State Office of the Spokesperson. https://www.state.gov/x-gender-marker-available-on-u-s-passports-starting-april-11/

Brown, E., & Perrett, D. I. (1993). What Gives a Face its Gender? *Perception*, *22*(7), 829–840. https://doi.org/10.1068/p220829

Burke, S. E., Jaurique, A., Valen, B. M., Wittlin, N. M., Ess, M., & LaFrance, M. (2023). *Forms of Psychological Bias Against Transgender Women and Men and People with Nonbinary Gender Identities*. https://doi.org/10.31234/osf.io/cpha2

Burke, S. E., & LaFrance, M. (2018). Perceptions of instability and choice across sexual orientation groups. *Group Processes & Intergroup Relations*, *21*(2), 257–279. https://doi.org/10.1177/1368430216663019

Burke, S. E., Perry, S. P., Dovidio, J. F., & LaFrance, M. (2023). Distinctive negative reactions to intermediate social groups. *Journal of Applied Social Psychology*, *53*(7), 600–630. https://doi.org/10.1111/jasp.12942

Buss, D. M. (1995). Psychological sex differences: Origins through sexual selection. *American Psychologist*, *50*(3), 164–168. https://doi.org/10.1037/0003-066X.50.3.164

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

Campbell, B., Schellenberg, E. G., & Senn, C. Y. (1997). Evaluating Measures of Contemporary Sexism. *Psychology of Women Quarterly*, *21*(1), 89–102. https://doi.org/10.1111/j.1471-6402.1997.tb00102.x

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words. *Psychological Science*, *32*(2), 218–240. https://doi.org/10.1177/0956797620963619

Chen, W., Ren, N., Young, A. W., & Liu, C. H. (2018). Interaction between social categories in the composite face paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(1), 34–49. https://doi.org/10.1037/xlm0000418

Chu, J., Nagpal, M., & Dobberfuhl, A. D. (2024). Utilization and Cost of Gender-affirming Surgery in the United States from 2012-2019. *Annals of Surgery*. https://doi.org/10.1097/SLA.0000000000006296

Clark, B. A., Veale, J. F., Townsend, M., Frohard-Dourlent, H., & Saewyc, E. (2018). Non-binary youth: Access to gender-affirming primary health care. *International Journal of Transgenderism*, *19*(2), 158–169. https://doi.org/10.1080/15532739.2017.1394954

Cohen, W., Maisner, R. S., Mansukhani, P. A., & Keith, J. (2020). Barriers To Finding A Gender Affirming Surgeon. *Aesthetic Plastic Surgery*, *44*(6), 2300–2307. https://doi.org/10.1007/s00266-020-01883-z

Coleman, E., Radix, A. E., Bouman, W. P., Brown, G. R., de Vries, A. L. C., Deutsch, M. B., Ettner, R., Fraser, L., Goodman, M., Green, J., Hancock, A. B., Johnson, T. W., Karasic, D. H., Knudson, G. A., Leibowitz, S. F., Meyer-Bahlburg, H. F. L., Monstrey, S. J., Motmans, J., Nahata, L., … Arcelus, J. (2022). Standards of Care for the Health of Transgender and Gender Diverse People, Version 8. *International Journal of Transgender Health*, *23*(sup1), S1–S259. https://doi.org/10.1080/26895269.2022.2100644

Connell, R. W., & Messerschmidt, J. W. (2005). Hegemonic Masculinity: Rethinking the Concept. *Gender & Society*, *19*(6), 829–859. https://doi.org/10.1177/0891243205278639

DeBruine, L., & Jones, B. (2021). *Face Research Lab London Set* (p. 281699214 Bytes) [Dataset]. figshare. https://doi.org/10.6084/M9.FIGSHARE.5047666.V5

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

dickey, lore m., Budge, S. L., Katz-Wise, S. L., & Garza, M. V. (2016). Health disparities in the transgender community: Exploring differences in insurance coverage. *Psychology of Sexual Orientation and Gender Diversity*, *3*(3), 275–282. https://doi.org/10.1037/sgd0000169

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Nature Communications*, *10*(1), 1258. https://doi.org/10.1038/s41467-019-09239-1

Downing, J., Holt, S. K., Cunetta, M., Gore, J. L., & Dy, G. W. (2022). Spending and Out-of-Pocket Costs for Genital Gender-Affirming Surgery in the US. *JAMA Surgery*, *157*(9), 799. https://doi.org/10.1001/jamasurg.2022.2606

Eagly, A. H. (1978). Sex differences in influenceability. *Psychological Bulletin*, *85*(1), 86–116. https://doi.org/10.1037/0033-2909.85.1.86

Eagly, A. H., & Carli, L. L. (2007). *Through the labyrinth: The truth about how women become leaders*. Harvard Business School Press.

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist*, *75*(3), 301–315. https://doi.org/10.1037/amp0000494

Eagly, A. H., & Steffen, V. J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. *Journal of Personality and Social Psychology*, *46*(4), 735–754. https://doi.org/10.1037/0022-3514.46.4.735

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

Elias, N. M. (2019). Transgender and Nonbinary Gender Policy in the Public Sector. In N. M. Elias, *Oxford Research Encyclopedia of Politics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.1168

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 *Advances in Experimental Social Psychology* (Vol. 23, pp. 1–74). Elsevier. https://doi.org/10.1016/S0065-2601(08)60317-2

Fiske, S. T., Neuberg, S. L., Beattie, A. E., & Milberg, S. J. (1987). Category-based and attribute-based reactions to others: Some informational conditions of stereotyping and individuating processes. *Journal of Experimental Social Psychology*, *23*(5), 399–427. https://doi.org/10.1016/0022-1031(87)90038-2

Flores, A. R., Haider-Markel, D. P., Lewis, D. C., Miller, P. R., Tadlock, B. L., & Taylor, J. K. (2018). Transgender prejudice reduction and opinions on transgender rights: Results from a mediation analysis on experimental data. *Research & Politics*, *5*(1), 205316801876494. https://doi.org/10.1177/2053168018764945

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–279. https://doi.org/10.1037/a0022327

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

Freeman, J. B., Stolier, R. M., & Brooks, J. A. (2020). Dynamic interactive theory as a domain-general account of social perception. In *Advances in Experimental Social Psychology* (Vol. 61, pp. 237–287). Elsevier. https://doi.org/10.1016/bs.aesp.2019.09.005

Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American Psychologist*, *56*(2), 109–118. https://doi.org/10.1037/0003-066X.56.2.109

Harmon, A. (2019, December). 'They' Is the Word of the Year, Merriam-Webster Says, Noting Its Singular Rise. *New York Times*. https://www.nytimes.com/2019/12/10/us/merriam-webster-they-word-year.html

Heilman, M. E. (2001). Description and Prescription: How Gender Stereotypes Prevent Women's Ascent Up the Organizational Ladder. *Journal of Social Issues*, *57*(4), 657–674. https://doi.org/10.1111/0022-4537.00234

Henry, P. J., & Sears, D. O. (2002). The Symbolic Racism 2000 Scale. *Political Psychology*, *23*(2), 253–283. https://doi.org/10.1111/0162-895X.00281

Herek, G. M. (2002). Heterosexuals' attitudes toward bisexual men and women in the United States. *The Journal of Sex Research*, *39*(4), 264–274. https://doi.org/10.1080/00224490209552150

Hernandez, J. (2023, April). Runners can identify as nonbinary in the Boston and London marathons. *National Public Radio*. https://www.npr.org/2022/09/14/1123035038/nonbinary-runners-boston-london-marathon

Houssayni, S., & Nilsen, K. (2018). Transgender Competent Provider: Identifying Transgender Health Needs, Health Disparities, and Health Coverage. *Kansas Journal of Medicine*, *11*(1), 1–18.

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

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

Jacques, S. A., Ross, D. E., & McCarty, M. K. (2022). Perceptions of Nonbinary Identifying Individuals: Through the Lens of Gender and Race. *Psi Chi Journal of Psychological Research*, *27*(1), 46–58. https://doi.org/10.24839/2325-7342.JN27.1.46

James, S., Herman, J. L., Rankin, S., Keisling, M., Mottet, L., & Anafi, M. A. (2016). *The report of the 2015 US transgender survey*. National Center for Transgender Equality.

Johnson, K. L., Iida, M., & Tassinary, L. G. (2012). Person (mis)perception: Functionally biased sex categorization of bodies. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1749), 4982–4989. https://doi.org/10.1098/rspb.2012.2060

Kattari, S. K., Atteberry-Ash, B., Kinney, M. K., Walls, N. E., & Kattari, L. (2019). One size does not fit all: Differential transgender health experiences. *Social Work in Health Care*, *58*(9), 899–917. https://doi.org/10.1080/00981389.2019.1677279

Kattari, S. K., & Hasche, L. (2016). Differences Across Age Groups in Transgender and Gender Non-Conforming People's Experiences of Health Care Discrimination, Harassment, and Victimization. *Journal of Aging and Health*, *28*(2), 285–306. https://doi.org/10.1177/0898264315590228

Kawakami, K., Amodio, D. M., & Hugenberg, K. (2017). Intergroup Perception and Cognition. In *Advances in Experimental Social Psychology* (Vol. 55, pp. 1–80). Elsevier. https://doi.org/10.1016/bs.aesp.2016.10.001

Kimura, D. (1992). Sex Differences in the Brain. *Scientific American*, *267*(3), 118–125. https://doi.org/10.1038/scientificamerican0992-118

Kranz, F., & Ishai, A. (2006). Face Perception Is Modulated by Sexual Preference. *Current Biology*, *16*(1), 63–68. https://doi.org/10.1016/j.cub.2005.10.070

Krysan, M. (2000). Prejudice, Politics, and Public Opinion: Understanding the Sources of Racial Policy Attitudes. *Annual Review of Sociology*, *26*(1), 135–168. https://doi.org/10.1146/annurev.soc.26.1.135

Lefevor, G. T., Boyd-Rogers, C. C., Sprague, B. M., & Janis, R. A. (2019). Health disparities between genderqueer, transgender, and cisgender individuals: An extension of minority stress theory. *Journal of Counseling Psychology*, *66*(4), 385–395. https://doi.org/10.1037/cou0000339

Lerner, J. E., & Robles, G. (2017). Perceived Barriers and Facilitators to Health Care Utilization in the United States for Transgender People: A Review of Recent Literature. *Journal of Health Care for the Poor and Underserved*, *28*(1), 127–152. https://doi.org/10.1353/hpu.2017.0014

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

Maccoby, E. E. (Ed.). (1966). *The development of sex differences* (1st ed.). Stanford Univ. Pr.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social Cognition: Thinking Categorically about Others. *Annual Review of Psychology*, *51*(1), 93–120. https://doi.org/10.1146/annurev.psych.51.1.93

Mao, J. M., Haupert, M. L., & Smith, E. R. (2019). How Gender Identity and Transgender Status Affect Perceptions of Attractiveness. *Social Psychological and Personality Science*, *10*(6), 811–822. https://doi.org/10.1177/1948550618783716

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

Masser, B., & Abrams, D. (1999). Contemporary Sexism: The Relationships Among Hostility, Benevolence, and Neosexism. *Psychology of Women Quarterly*, *23*(3), 503–517. https://doi.org/10.1111/j.1471-6402.1999.tb00378.x

Matsuno, E., & Budge, S. L. (2017). Non-binary/Genderqueer Identities: A Critical Review of the Literature. *Current Sexual Health Reports*, *9*(3), 116–120. https://doi.org/10.1007/s11930-017-0111-8

Miller, P. R., Flores, A. R., Haider-Markel, D. P., Lewis, D. C., Tadlock, B., & Taylor, J. K. (2020). The Politics of Being "Cait": Caitlyn Jenner, Transphobia, and Parasocial Contact Effects on Transgender-Related Political Attitudes. *American Politics Research*, *48*(5), 622–634. https://doi.org/10.1177/1532673X20906460

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

Morgenroth, T., Axt, J. R., & Westgate, E. C. (2022). What Underlies the Opposition to Trans-Inclusive Policies? The Role of Concerns About Male Violence Versus Attitudes Toward Trans People. *Personality and Social Psychology Bulletin*, 014616722211372. https://doi.org/10.1177/01461672221137201

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., McKee, K., Huff, D., & Abdi, H. (1998). The perception of face gender: The role of stimulus structure in recognition and classification. *Memory & Cognition*, *26*(1), 146–160. https://doi.org/10.3758/BF03211378

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

Perez-Arche, H., & Miller, D. J. (2021). What Predicts Attitudes toward Transgender and Nonbinary People? An Exploration of Gender, Authoritarianism, Social Dominance, and Gender Ideology. *Sex Roles*, *85*(3–4), 172–189. https://doi.org/10.1007/s11199-020-01212-3

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*(2, Pt.1), 304–308. https://doi.org/10.1037/h0028558

Pratto, F., Stallworth, L. M., & Conway‐Lanz, S. (1998). Social Dominance Orientation and the Ideological Legitimization of Social Policy [1]. *Journal of Applied Social Psychology*, *28*(20), 1853–1875. https://doi.org/10.1111/j.1559-1816.1998.tb01349.x

Puckett, J. A., Cleary, P., Rossman, K., Mustanski, B., & Newcomb, M. E. (2018). Barriers to Gender-Affirming Care for Transgender and Gender Nonconforming Individuals. *Sexuality Research and Social Policy*, *15*(1), 48–59. https://doi.org/10.1007/s13178-017-0295-8

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

Richards, C., Bouman, W. P., & Barker, M.-J. (Eds.). (2017). *Genderqueer and Non-Binary Genders*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-51053-2

Ridgeway, C. L. (2001). Gender, Status, and Leadership. *Journal of Social Issues*, *57*(4), 637–655. https://doi.org/10.1111/0022-4537.00233

Roberts, T., & Bruce, V. (1988). Feature Saliency in Judging the Sex and Familiarity of Faces. *Perception*, *17*(4), 475–481. https://doi.org/10.1068/p170475

Rubin, J. D., Atwood, S., & Olson, K. R. (2020). Studying Gender Diversity. *Trends in Cognitive Sciences*, *24*(3), 163–165. https://doi.org/10.1016/j.tics.2019.12.011

Rudman, L. A., & Glick, P. (2001). Prescriptive Gender Stereotypes and Backlash Toward Agentic Women. *Journal of Social Issues*, *57*(4), 743–762. https://doi.org/10.1111/0022-4537.00239

Sanchez, D. T., & Bonam, C. M. (2009). To Disclose or Not to Disclose Biracial Identity: The Effect of Biracial Disclosure on Perceiver Evaluations and Target Responses. *Journal of Social Issues*, *65*(1), 129–149. https://doi.org/10.1111/j.1540-4560.2008.01591.x

Schmader, T. (2023). Gender Inclusion and Fit in STEM. *Annual Review of Psychology*, *74*(1), 219–243. https://doi.org/10.1146/annurev-psych-032720-043052

Schudson, Z. C., & Morgenroth, T. (2022). Non-binary gender/sex identities. *Current Opinion in Psychology*, *48*, 101499. https://doi.org/10.1016/j.copsyc.2022.101499

Sevelius, J., & Jenness, V. (2017). Challenges and opportunities for gender-affirming healthcare for transgender women in prison. *International Journal of Prisoner Health*, *13*(1), 32–40. https://doi.org/10.1108/IJPH-08-2016-0046

Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A. (1995). Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology*, *68*(2), 199–214. https://doi.org/10.1037/0022-3514.68.2.199

Swim, J. K., & Sanna, L. J. (1996). He's Skilled, She's Lucky: A Meta-Analysis of Observers' Attributions for Women's and Men's Successes and Failures. *Personality and Social Psychology Bulletin*, *22*(5), 507–519. https://doi.org/10.1177/0146167296225008

Tate, C. C., Ben Hagai, E., & Crosby, F. J. (2020). *Undoing the Gender Binary* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108584234

Tate, C. C., Youssef, C. P., & Bettergarcia, J. N. (2014). Integrating the Study of Transgender Spectrum and Cisgender Experiences of Self-Categorization from a Personality Perspective. *Review of General Psychology*, *18*(4), 302–312. https://doi.org/10.1037/gpr0000019

Taylor, A. J., & Parcel, T. L. (2019). Proximity and the principle-policy gap in white racial attitudes: Insight from views of student assignment policies in Wake County, North Carolina. *Social Science Research*, *78*, 95–103. https://doi.org/10.1016/j.ssresearch.2018.12.014

Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, *36*(7), 778–793. https://doi.org/10.1037/0022-3514.36.7.778

The Trevor Project. (2021). *The Trevor Project Research Brief: Diversity of Nonbinary Youth* [Research Brief]. https://www.thetrevorproject.org/research-briefs/diversity-of-nonbinary-youth/

Timothy Cavanaugh, Ruben Hopwood, & Cei Lambert. (2016). Informed Consent in the Medical Care of Transgender and Gender-Nonconforming Patients. *AMA Journal of Ethics*, *18*(11), 1147–1155. https://doi.org/10.1001/journalofethics.2016.18.11.sect1-1611

Tuch, S. A., & Hughes, M. (2011). Whites' Racial Policy Attitudes in the Twenty-First Century: The Continuing Significance of Racial Resentment. *The ANNALS of the American Academy of Political and Social Science*, *634*(1), 134–152. https://doi.org/10.1177/0002716210390288

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

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

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

Wilson, B. D. M., & Meyer, I. H. (2021). *Nonbinary LGBTQ Adults in the United States* [Brief]. Williams Institute, UCLA School of Law.

https://williamsinstitute.law.ucla.edu/wp-content/uploads/Nonbinary-LGBTQ-Adults-Jun-2021.pdf

Worthen, M. G. F. (2013). An Argument for Separate Analyses of Attitudes Toward Lesbian, Gay, Bisexual Men, Bisexual Women, MtF and FtM Transgender Individuals. *Sex Roles*, *68*(11–12), 703–723. https://doi.org/10.1007/s11199-012-0155-1

Yang, X., & Dunham, Y. (2019). Hard to disrupt: Categorization and enumeration by gender and race from mixed displays. *Journal of Experimental Social Psychology*, *85*, 103893. https://doi.org/10.1016/j.jesp.2019.103893

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

Zhao, M., & Hayward, W. G. (2010). Holistic processing underlies gender judgments of faces. *Attention, Perception, & Psychophysics*, *72*(3), 591–596. https://doi.org/10.3758/APP.72.3.591

**Flexibility in continuous judgements of gender/sex and race**

S. Atwood[1,*^], Dominic J. Gibson[2,*], Sofía Briones Ramírez[1], and Kristina R. Olson[1,2]

[1]Department of Psychology, Princeton University, Princeton, NJ 08540 USA.

[2]Department of Psychology, University of Washington, Seattle, WA 98105 USA.

*These authors contributed equally.

^Correspondence to: satwood@princeton.edu

**Word Count: 12373**

**Author Note:** 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. All data, study materials, code, and Supplementary Material are available at the project page on the Open Science Framework (https://osf.io/3xzq9/?view_only=da219dd6cc8b4a548e42ab74031098ec). 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. Additionally, we declare no financial conflicts or competing interests that arise from this research or its publication.
**Dissertation Note:** This chapter is currently in published at *JEP:G*. I share first authorship with Dominic J. Gibson on this work.

**Abstract**

Across six pre-registered studies ($N$ = 1292; 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 pre-existing 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.

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

## Statement of Public Significance

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.

How do people think about and use social dimensions like gender/sex[2] 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 MacLin & Malpass, 2001; Levin & Angelone, 2002; Walker & Tanaka, 2003; for gender/sex see Campanella et al., 2001; Freeman, Rule, et al., 2010). For instance, when faces are morphed such that a set of stimuli are created that transition gradually from a male to female face, participants' categorization of those faces abruptly shifts 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.

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

To the extent that people's ideas about social dimensions are biased towards 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, even as people regularly 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, able to be shifted with specific experiences?

In the current work, we sought insight into these questions by asking participants to make judgments[3] about gender/sex (our primary focus) and race along a continuous spectrum and comparing the patterns of responses to how participants make judgments about non-social dimensions, color and number (Studies 1 & 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

[3] 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 Methods section in Study 1a) .

identity/extended experience (Study 3) and/or their immediate experience in the context of the experiment (Studies 4 & 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. Olian et al., 1988; Bertrand & Mullainathan, 2004; 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. Ridgeway, 2011; Allport, 1954).

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 (Eagly & Wood, 1999; Bussey, 2011). Infants differentiate members of gender/sex groups as early as 3 months old (Quinn et al., 2002). Moreover, between 3-5 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 non-social 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 feature). This multidimensionality may mean that social dimensions are more intrinsically and even arbitrarily defined by discrete categories.

**Categorical perception of gender/sex and race**[4]

Due to these or other factors, people appear to rely on discrete categories even when sorting faces that vary continuously (Levin, 1996; Levin & Beale, 2000; Levin & Angelone, 2002; 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 selected the male vs. female options? Or did half of 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 et al., 2014; Chen & Hamilton, 2012) and have poorer memory for faces with ambiguous races (Pauker et al., 2009), suggesting uncertainty and a lack of attention around

[4] There are ongoing debates about whether social perception research of the type discussed in this paper 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.

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 (2010) 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". 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 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 towards the ends of social dimensions like race or gender/sex than faces that fall in the middle and may 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 social 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; Ma et al., 2011; Malpass & Kravitz, 1969; Meissner & Brigham, 2001). 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; Fallshore & Schooler, 1995; Correll et al., 2021; 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, 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; Meissner & Brigham, 2001; Lee et al., 2011).

The own-race advantage has also been shown to extend to members of "intermediary" racial groups. 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 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, 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 that assessed gender/sex and race face judgments 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 & Benedict, 2021, Cafe Dataset; LoBue & Thrasher, 2015), with a few notable exceptions (e.g. Freeman 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 non-social stimuli along non-social 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 non-social 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 OSF project page (see https://osf.io/3xzq9/?view_only=da219dd6cc8b4a548e42ab74031098ec). 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/?view_only=adacd4e244cb40dbac00f94466970488). Often the main paper 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 Supplementary Material. Importantly, the findings highlighted in the main paper 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 University of Washington and Princeton University. These preregistrations include all exclusion criteria used

in the current manuscript as well as additional analyses that go beyond the scope of the present paper. However, all preregistered analyses are fully reported in the manuscript or the Supplementary Material (see https://osf.io/d3htw/?view_only=5e33fa9f9270497982f25d75a3a9b3eb).

## 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 main manuscript 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 4000ms response period or responded in 300ms or under (a sign of careless responding), as per our preregistration (https://osf.io/u4dxw/?view_only=adacd4e244cb40dbac00f94466970488). 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 number trials, so no participants were dropped by that criterion. However, we dropped data from color trials of 1 participant who identified as color blind, as per our preregistration (https://osf.io/7b6qd/?view_only=adacd4e244cb40dbac00f94466970488). Finally, we excluded 3.4% of all responses based on response time.

**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 | | | | | | |
| Mean (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 |

Table 1. Demographic information for all studies.

***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 stimuli 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, 2021), the Transphobia Scale (Nagoshi & Nagoshi, 2008), and a pair of feelings thermometers towards cisgender and transgender people, respectively (Alwin, 1997; Dunton & Fazio, 1997; IJzerman & Semin, 2009) (see Supplementary Material pgs. 4-5 for reporting on all survey measures).

***Task Design.*** All participants completed 3 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 9 faces generated by a single male/female pair were always run in the same block) to ensure our observed results were not design-dependent.

***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 faces in each of 3 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 9 distinct morphs from each of the 36 pairs of faces (e.g., 10% of the female face and 90% of the male, etc; see Fig. 1 below for an example set of face gender/sex stimuli and pgs. 1-2 in the Supplementary Material for additional details about our face morphing procedure for all face stimuli in this paper). At test, participants in Study 1a sorted all morphs from half the face pairs of each race (162 faces total), and morphs from the other half of the face pairs were given to a different set of participants in Study 1b (see https://osf.io/u69yj/?view_only=c9374011248b42fcb1c257242a775b23 for full sets of face stimuli for both studies). This meant that Study 1b was a replication with unique stimuli (in line for calls for 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 6 distinct pairs of endpoints. Each of the 6 colors was represented twice to create 6 total continua, and we purposefully avoided pairing complementary colors due to their propensity to cancel

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 9 intermediary colors that were equidistant from one another in RGB color space for each of the 6 pairs (see Fig 1 for an example set of color stimuli). At test, participants sorted all 54 color stimuli in both Studies 1a and 1b (see https://osf.io/u69yj/?view_only=c9374011248b42fcb1c257242a775b23 for a full set of color stimuli).

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

***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 Supplementary Material on pgs. 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 0-100 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.

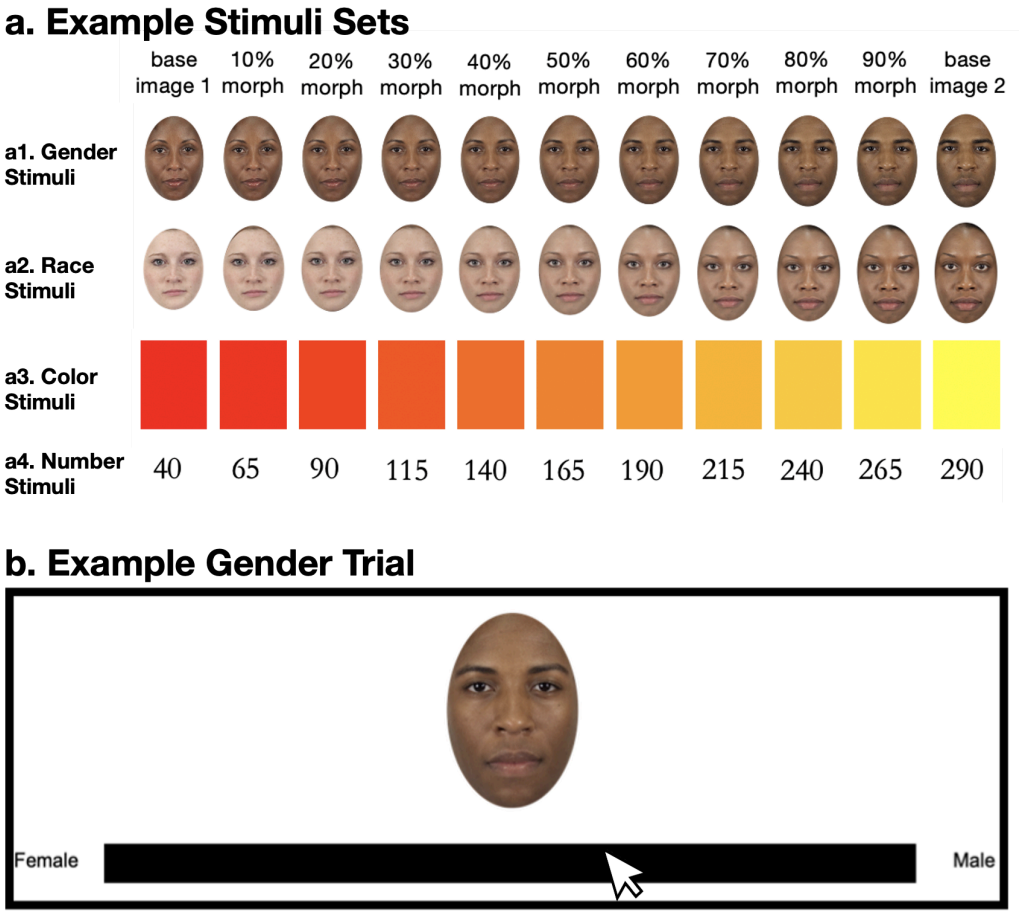



Figure 1. Panel A contains example stimuli in each domain presented throughout the paper. The outermost stimuli were used as base images but were never seen by participants, while all 9 inner

morphs were presented to participants in a random order (along with other stimuli in the same category). In Studies 3-4, 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.

***Overview of Statistical Approach.*** In the six studies presented in this paper, 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 (that is, 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 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[5] 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 paper are presented within 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

[5] Note: 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.

understand patterns of accuracy across stimuli, we visually inspect and describe sorting patterns at different levels of stimulus ambiguity across all 3 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 Supplementary Material (see pgs. 3-4 (Study 1a), pg. 5 (Study 1b), pg. 8 (Study 2), pg. 22 (Study 3), and pg. 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 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 paper, we report analyses run in R Studio (Version 2023.03.1+446) and use scaled down values of error, intermediateness, subjective judgements, 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).[6]

[6] 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 Supplementary Material.

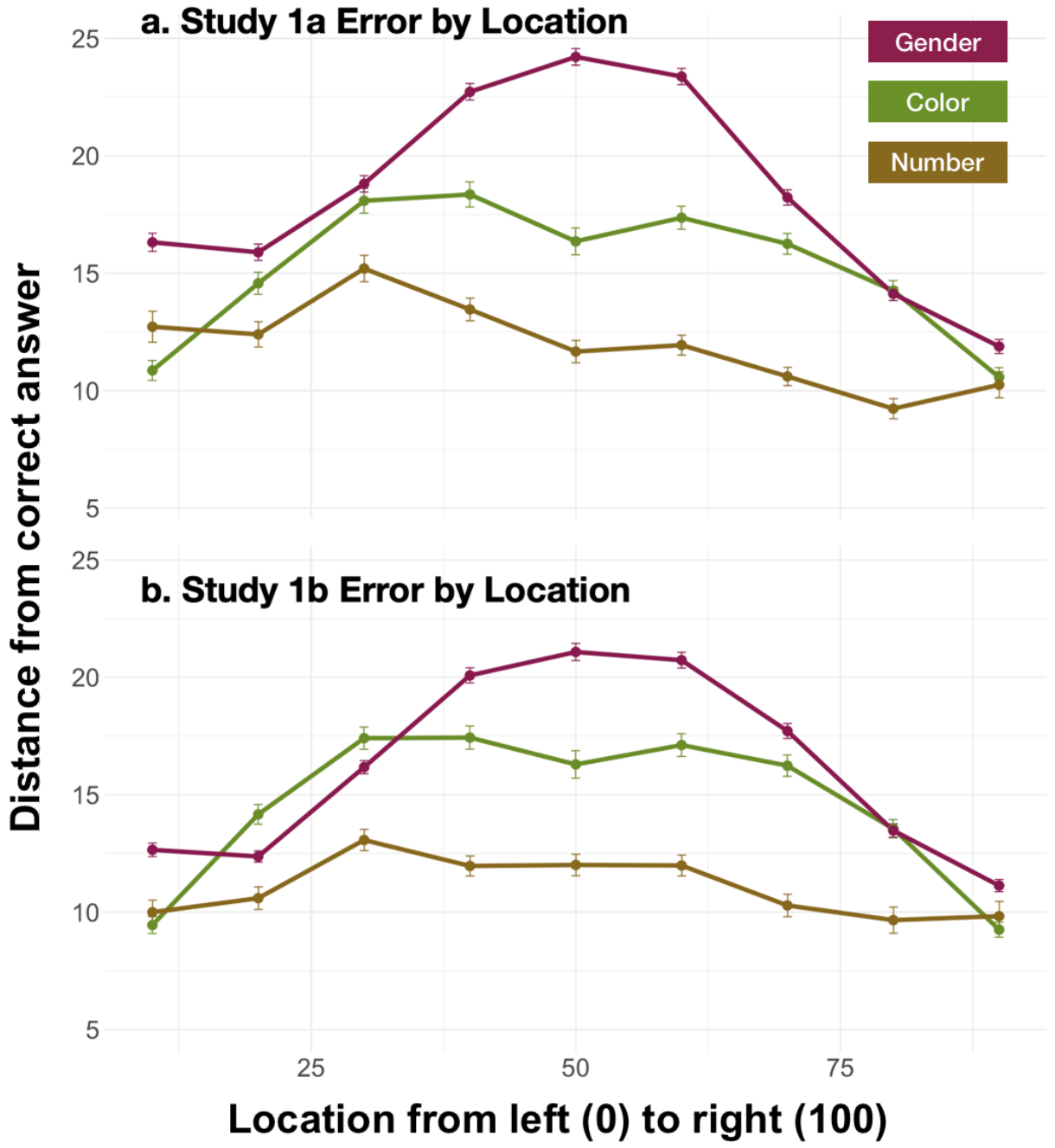



Figure 2. Average distance from correct answer (or error) by objective (correct) location and by domain in Studies 1a and 1b. Error bars represent standard error.

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' judgements were distributed across the entire range of possible values. Further, their error in judgements was highest for these intermediate stimuli (Figure 2).

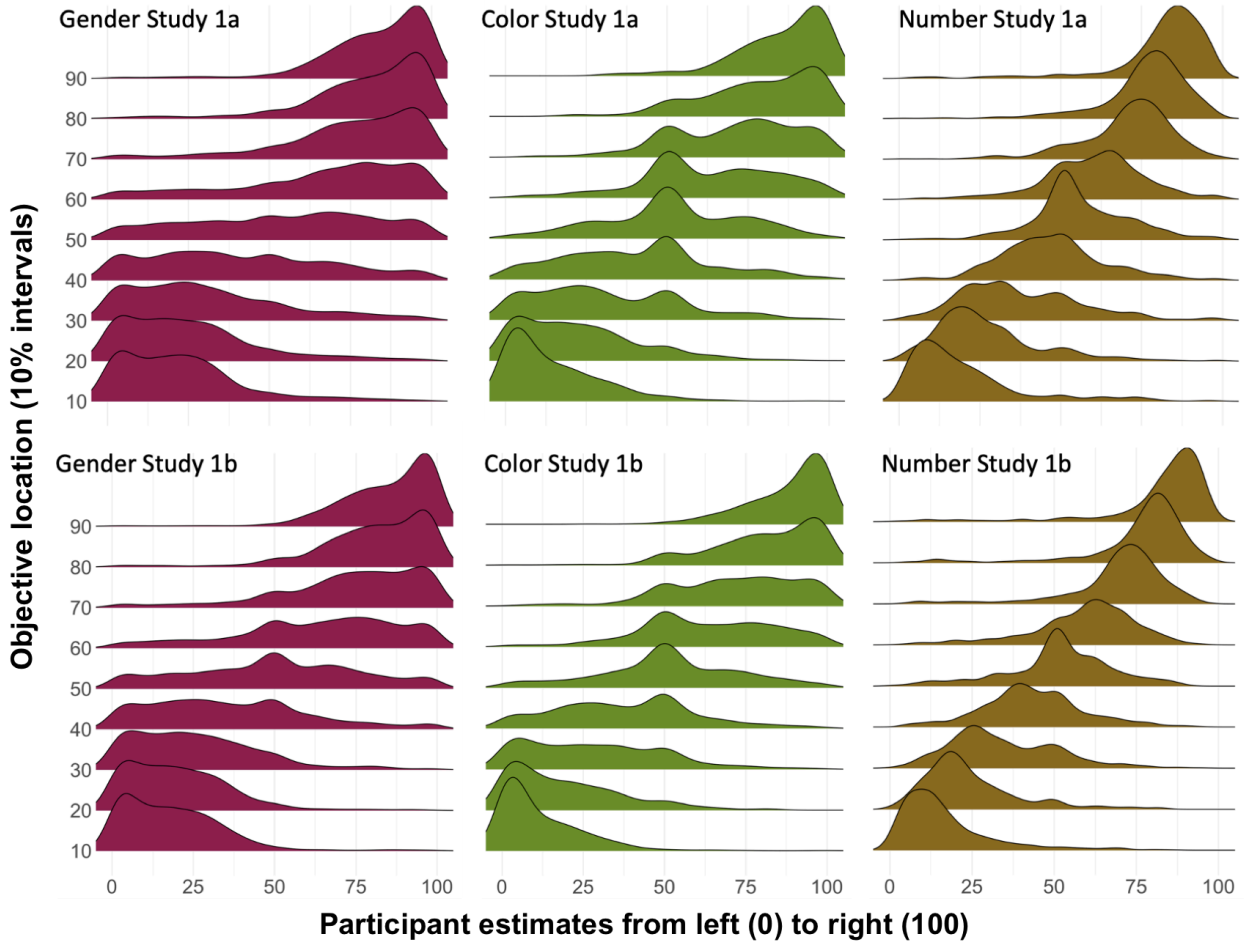



Figure 3. Distribution of participant estimates of stimulus location along the continuum by objective (correct) location in Studies 1a and 1b.

In addition to our visual inspection, we found evidence of both linear and a non-linear 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 (Studies 1a; $b$ = .118, standard error (*S.E.*) = .004, $p < .001$ and 1b; $b$ = .128, *S.E.* = .003, $p < .001$), participants judgments indicated significant cubic (Study 1a; $b$ = -.003, *S.E.* = .0002, $p < .001$; Study 1b; $b$ = -.003, *S.E.* = .0001, $p < .001$) components, suggesting a non-linear relationship between objective and subjective gender/sex judgments (see Figure 2). Similarly, for numbers, participants' judgements indicated significant linear (Study 1a; $b$ = .008, *S.E.* = .003, $p < .001$; Study 1b; $b$ = .086, *S.E.* = .002, $p < .001$), and quadratic (Study 1a only; $b$ = -.004, *S.E.* = .001, $p$ = .002) components. For colors, participants' judgments suggested a significant linear component (Study 1a; $b$ = .103, *S.E.* = .003, $p < .001$; Study 1b; $b$ = .114, *S.E.* = .003, $p < .001$) as well as a cubic component (Study 1a; $b$ = -.001, *S.E.* = .0002, $p <.001$; Study 1b; $b$ = -.001, *S.E.* = .0002, $p < .001$). Collectively, these results suggest that there both linear and non-linear 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.

***Discussion***

Across two studies that varied several aspects of study design, we consistently found evidence of both linear and non-linear 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 yet they 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 if 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 2 blocks of race stimuli (which contained equal numbers of faces that were male and female) and 2 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 & 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 Supplementary Material pgs. 7-8 for reporting on survey measures).

***Task Design.*** As in Study 1b, stimuli were grouped together based on the base pair used to create them. Additionally, we counterbalanced the location of labels (which were male and female for gender/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 4 same-race face pairs from each of 3 racial groups (Black, East Asian[7], and White) morphed at a 10% gradient for a total of 108 faces that participants sorted at test.

[7] As opposed to West Asian as denoted in the London Face Database

Race stimuli were created by selecting the 4 male faces and 4 female faces that were most consistently rated as Asian, Black, and White (for a total of 12 male faces and 12 female faces, evenly spread across 3 races) from the Chicago Face Database (Ma, et al., 2015). Then, we semi-randomly paired faces across race (i.e. Asian-White, Black-Asian, and White-Black) to end up with 2 male and 2 female pairs for each cross-race combination. The resulting 12 face pairs served as the endpoints, which were systematically morphed at a 10% gradient to create 9 distinct morphs (see Fig. 1 below for an example set of race stimuli). At test, participants sorted all morphs (but not original faces) for a total of 108 faces (see https://osf.io/u69yj/?view_only=c9374011248b42fcb1c257242a775b23 for full set of race stimuli).

***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 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) (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 judgements of race (see Figure 4, Panel A).

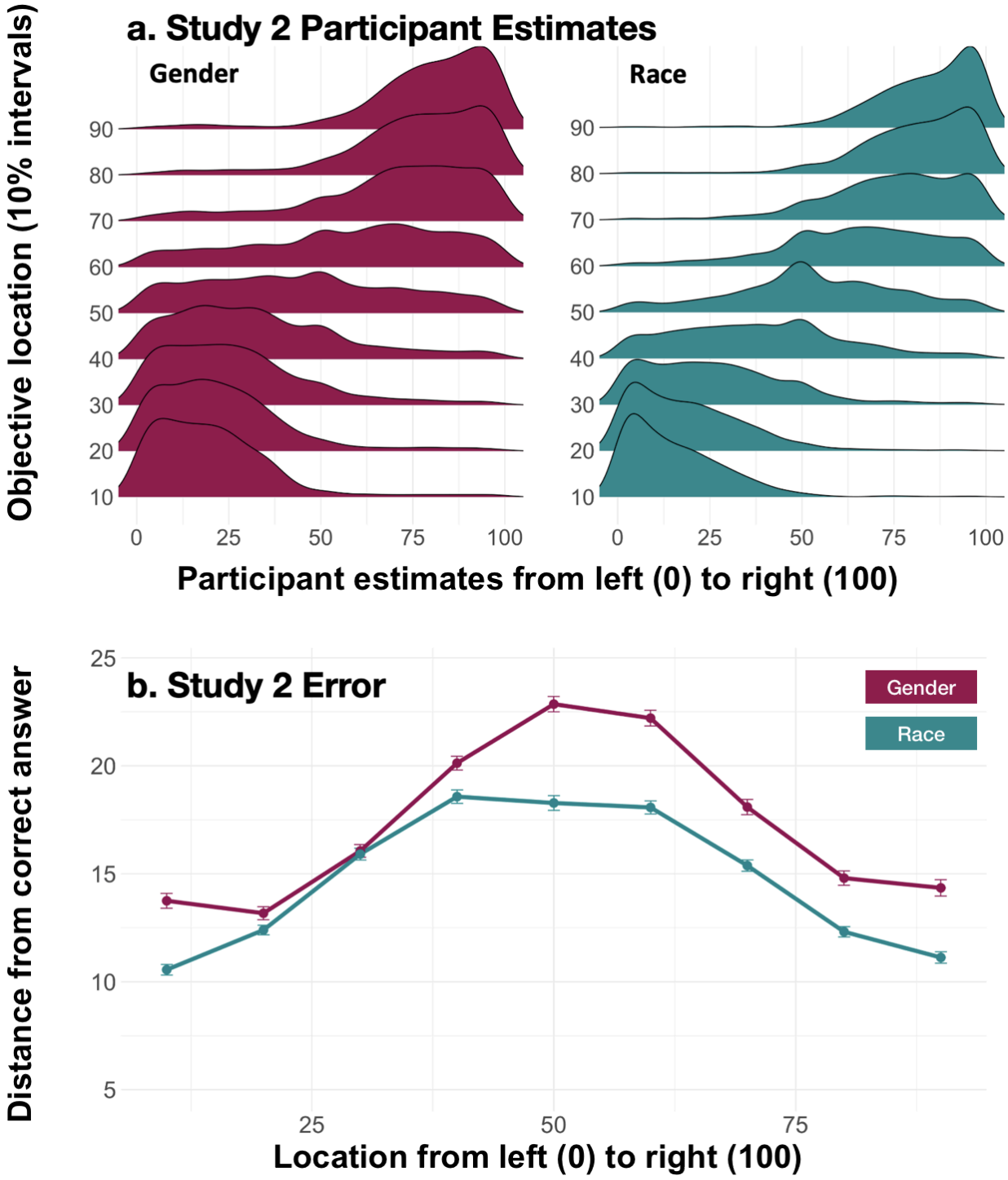


Figure 4. Panel A is the distribution of participant estimates of stimulus location by objective location in Study 2 and Panel B is average distance from correct answer by objective location and by domain in Study 2. Error bars represent standard error.

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, *S.E.* = .003, $p < .001$), quadratic ($b$ = .008, *S.E.* = .001, $p < .001$), cubic ($b$ = -.003, *S.E.* = .0001, $p < .001$), and quartic ($b$ = -.0003, *S.E.* = .0001, $p < .001$) components for gender/sex, and race (linear; $b$ = .131, *S.E.* = .003, $p < .001$, quadratic; $b$ = -.005, *S.E.* = .0009, $p < .001$, cubic; $b$ = -.003, *S.E.* = .0001, $p < .001$, quartic; $b$ = .0002, *S.E.* = .00005, $p < .001$) suggesting a strong linear relationship and some evidence of non-linear relationships between objective location and subjective judgment in both gender/sex and race judgements (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 non-linear 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.

## 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 (Nagoshi et al., 2012; Iantaffi & Bockting, 2011). 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; de Vries, 2012; Bradford et al, 2019) 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; Gagné et al., 1997; Factor & Rothblum, 2008; 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 in a more rushed way

(within 2000 ms[8]) and at the same speed as in the previous studies (4000ms). This manipulation was selected in anticipation of a group difference between the gender diverse and cisgender heterosexual participants at 4000ms. 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 rushed. 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 paper, 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 1 in the Supplementary Material (pgs. 10-11). Similar to prior studies, we excluded data from trials where participants responded in 300ms or less and data from participants who indicated face blindness. In this study we also preregistered to excluded anyone who exceeded the maximum amount of time allotted to sort each face (2000ms in the "fast" condition and 4000ms 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 9 participants (not included in the total above).

[8] Chosen based on the results of a pilot test run during summer of 2021; for more information see Supplementary Material pg. 9.

***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 2000ms (in the "fast" condition) or 4000ms (in the "slow" condition) to make each judgment. In order 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 6 same-race face pairs from each of 3 racial groups (Black, East Asian, and White) morphed at a 10% gradient for a total of 108 faces that participants sorted at test. As in previous studies, stimuli were grouped together based on the base pair used to create them and 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 4 of the most phenotypically masculine and 4

of the most phenotypically feminine faces within Asian, Black and White races from the Chicago Face Database (Ma et al., 2015) morphed at a 10% gradient.

***Results***

We ran a multilevel regression model predicting error from intermediateness, participant group, response window, and their 2 and 3-way interactions (see Supplementary Material pg. 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 was 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)[9].

[9] In the preregistration, our two primary groups of interest were cisgender heterosexuals and gender diverse participants. However, we also pre-registered 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 Supplementary Material starting on pg. 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.

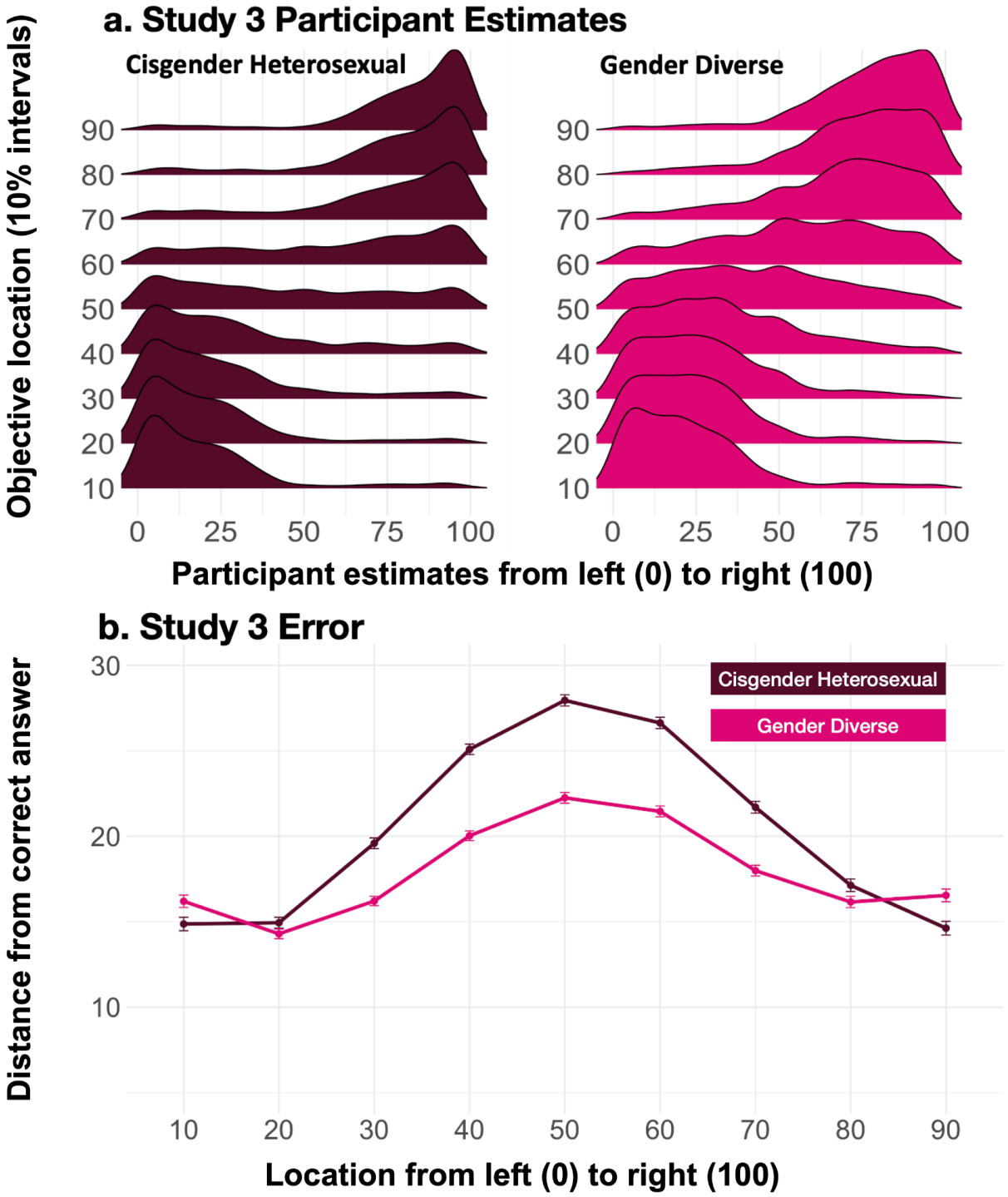



Figure 5. Panel A is the distribution of participant estimates of stimulus location by objective location in Study 3 and Panel B is average distance from correct answer by objective location and by domain in Study 3. Error bars represent standard error.

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 $p$'s > .1; see Figure 5; Table 2 in Supplementary Material). 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 2 in Supplementary Material). Unlike androgyny, we found no meaningful differences in the error of participants' judgements based on the amount of time they were given (2000ms vs. 4000ms) to sort each face (see Supplementary Material pgs. 14-15 for more extensive discussion).

***Discussion***

We found that gender diverse participants judged 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, or gender transitions, or differences in motivation cannot be determined by this study. However, the following studies aim to ask whether people who lack the kinds of long-term experience of gender diverse people might still be capable of a change in 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 for intermediary faces.

### *Method*

***Participants and Exclusions***. In Study 4, a total of 228 individuals (randomly assigned to either the 2-category or 3-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 pre-registration, 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 2-category label (similar to past studies) or 3-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 2 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 androgynous case; therefore all labels were converted to symbols. Participants in the 3 category condition also saw a third set of images and an additional label, which represented an intermediate category which 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 Supplementary Material Figure S1, Panels A and B). Finally, participants were given one last reminder that displayed all labels (as symbols) and associated faces they learned about mapped out on the continuum (See Supplementary Material Figure S1, Panel C)[10]. 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 3-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, one containing race stimuli and the other gender/sex stimuli in random order each containing 76 trials.

**Task Design.** As in Studies 1b, 2, and 3, stimuli in Study 4 were grouped together based on the base 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

[10] 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 judgements 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.

were asked to fill out a short survey following 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 were 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 4 out of the 12 morphs for both race and gender/sex, resulting in 76 gender/sex stimuli and 76 race stimuli sorted at test.

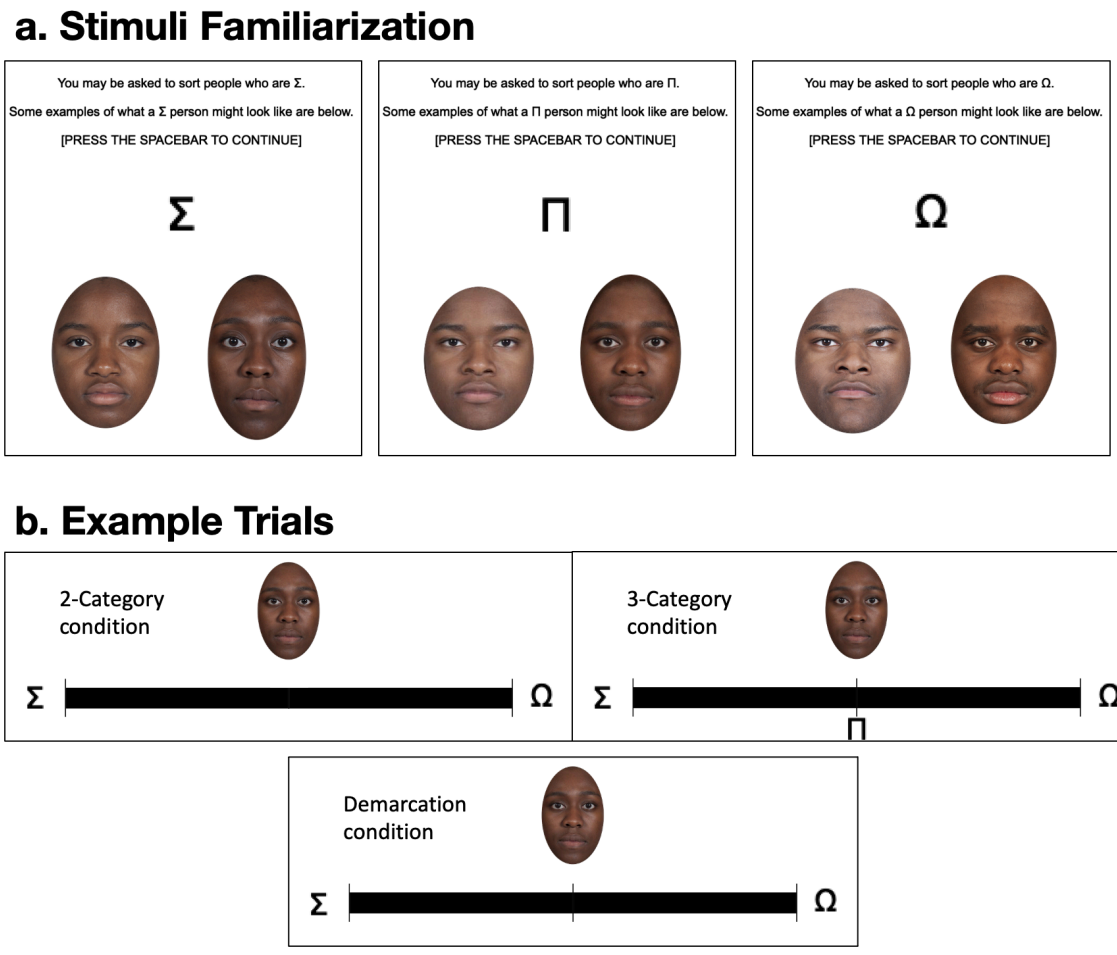


Figure 6. Panel A: Symbol familiarization in the 3-category gender/sex condition; Panel B: Example 2 and 3-category task trials (as seen in Studies 4 & 5) and demarcation task trial (as seen in Study 5).

***Results***

As in previous studies, we conducted a multilevel regression model predicting error from intermediateness, condition, domain, and their 2-and 3-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 3 category condition. We found that participants demonstrated greater error when perceiving more intermediary stimuli ($b$ = .100, $t(265) = 8.34$, $p < .001$; see Figure 8). Participants in the 3-category condition demonstrated greater overall error than participants in the 2-category condition ($b$ = .032, $t(238) = 3.24$, $p = .001$) . Critically, participants in the 3-category condition showed less error for intermediate faces than those in the 2-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 3-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 2-category condition than those in the 3-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 7, 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.

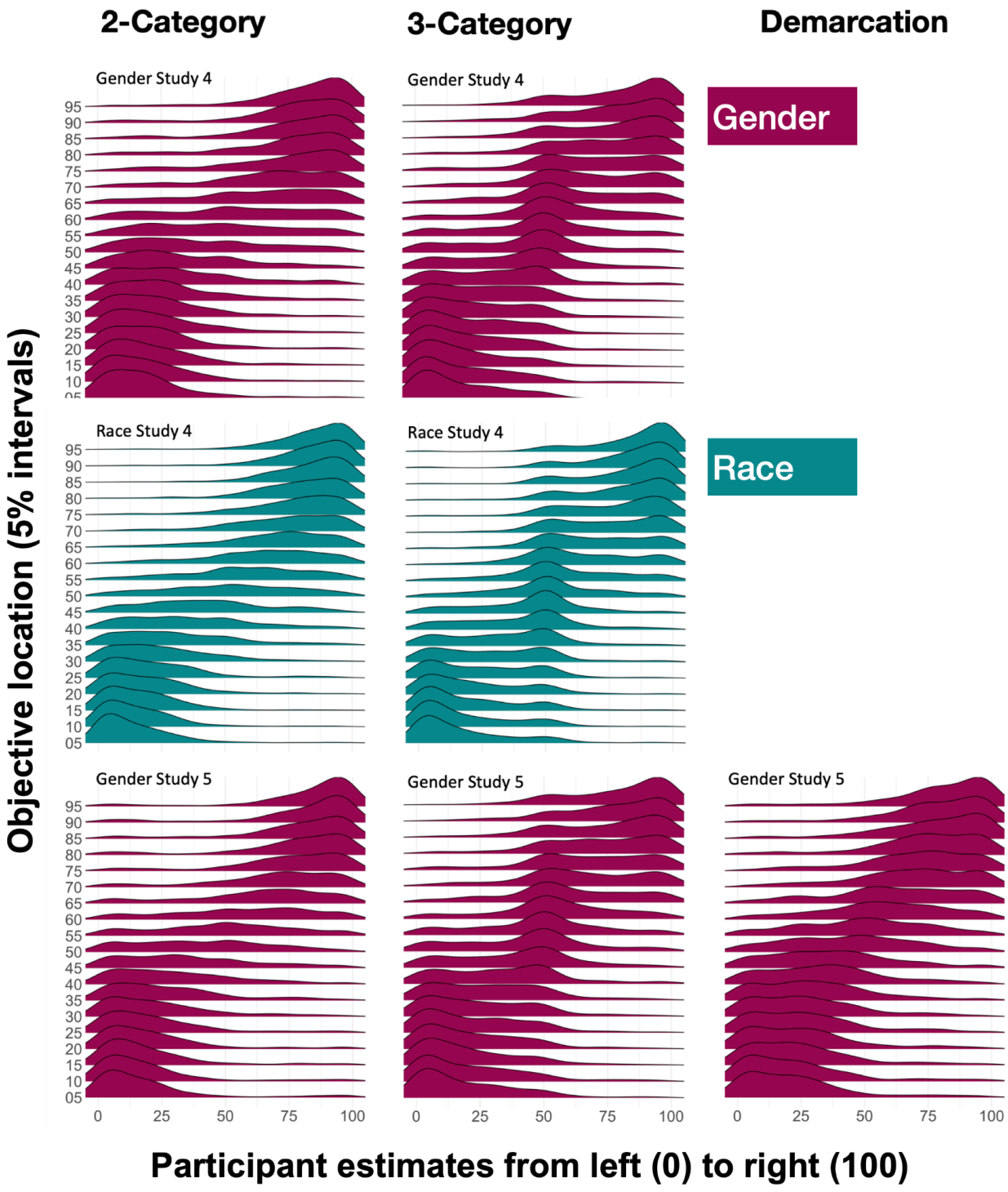



Figure 7. Distribution of participant estimates of stimulus location by objective location in Studies 4 and 5


a. Study 4 Gender Error by Location
b. Study 4 Race Error by Location
c. Study 5 Error by Location
Distance from correct answer
Location from left (0) to right (100)
25
20
15
10
5
25
50
75
Gender 2-Category
Gender 3-Category
Race 2-Category
Race 3-Category
Gender Demarcation

Figure 8. Average distance from correct answer by objective location and by domain and condition in Studies 4 and 5. Error bars represent standard error.

***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 3-category condition increased the saliency of the center of the continuum or that it created an actual third category in the minds of our participants. In Study 5, we compare three conditions, the 2- and 3-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 3-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 3-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 2-category condition; $n$ = 112, 3-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 pre-registration, 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 2-category label and 3-category label version of the task, we introduced a third condition that was identical to the 2-category condition except that participants in this condition sorted stimuli on a black line with a demarcation (but no label) at the midpoint (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 2-category condition), and then saw two example faces sorted at 25% and 75% of the way between the endpoints of the continuum. At test, participants saw a black line with 2 labels (2-category condition), 3 labels (3-category condition), or 2 labels plus a demarcation in the middle (demarcation condition), and were asked to sort each stimulus along the line in relation to the labels at either end (see Figure 6, Panel B). In 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 completion of the task.

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

***Results***

In order 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 2-category condition than 3-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 2-category condition ($b$ = -.032, $t$(320) = -2.07, $p$ = .039), but more error than those in the 3-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 3-category condition, $t$(825) = 4.00, $p$ < .001, 99% CI = [.00617, .07515], but no different than those in the 2-category condition, $t$(858) = -2.00, $p$ > .05, 99% CI = [-.0517, .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, our work and past work (Campanella et al., 2001) is consistent in suggesting 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% 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's 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 sub-categories, 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 multi-racial 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 builds 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 (Peery & Bodenhausen, 2009; Chen & Hamilton, 2012; Gaither et al., 2019; Young et al., 2021; Chen et al., 2023; Levy et al., 2023). 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 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, and warmth 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

Supplementary Material pgs. 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 would be 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 Supplementary Material pgs. 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 (versus two) (Paas & Van Merriënboer, 1994; Sweller, 1988). However, unlike Chen & 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 tri-modal 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, for example, 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 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 may be limited 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 non-facial features such as clothing, voice, and gait 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 in this sense. 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 Pub. Co.

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

Atwood, S., Gibson, D. J., Ramírez, S. B., & Olson, K. R. (n.d.). *Flexibility in gender/sex and race perception materials and data* [Repository]. Open Science Framework. https://osf.io/3xzq9/?view_only=da219dd6cc8b4a548e42ab74031098ec

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*. 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. *The American Economic Review*, *94*(4), 991–1013.

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. *The 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 Faces1. *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 New York. 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). Gender recognition from body. *Proceedings of the 16th ACM International Conference on Multimedia*, 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*. 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*, 104058. https://doi.org/10.1016/j.jesp.2020.104058

Crisp, R. J., & Hewstone, M. (2007). Multiple Social Categorization. In *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), e0141353. https://doi.org/10.1371/journal.pone.0141353

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

de Heering, A., & Rossion, B. (2008). Prolonged Visual Experience in Adulthood Modulates Holistic Face Perception. *PLoS ONE*, *3*(5), e2317. https://doi.org/10.1371/journal.pone.0002317

DeBruine, L. (2018). *debruine/webmorph: Beta release 2* (v0.0.0.9001) [Computer software].

Zenodo. https://doi.org/10.5281/ZENODO.1162670

DeBruine, L., & Jones, B. (2021). *Face Research Lab London Set* (p. 281699214 Bytes) [dataset]. figshare. https://doi.org/10.6084/M9.FIGSHARE.5047666.V5

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 *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 *Proceedings of the eighteenth annual conference of the Cognitive Science Society* (p. 6). Routledge.

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The *R* Package **geepack** for Generalized Estimating Equations. *Journal of Statistical Software*, *15*(2). https://doi.org/10.18637/jss.v015.i02

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

Kahn, T., Bosch, J., Levitt, M. F., & Goldstein, M. H. (1975). Effect of sodium nitrate loading on electrolyte transport by the renal tubule. *The American Journal of Physiology*, *229*(3), 746–753. https://doi.org/10.1152/ajplegacy.1975.229.3.746

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).

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*.

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

Mallett, R. K., & Wagner, D. E. (2011). The unexpectedly positive consequences of confronting sexism. *Journal of Experimental Social Psychology*, *47*(1), 215–220. https://doi.org/10.1016/j.jesp.2010.10.001

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

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].

Morgenroth, T., Van Der Toorn, J., Pliskin, R., & McMahon, C. E. (2023). Gender Nonconformity Leads to Identity Denial for Cisgender and Transgender Individuals. *Social Psychological and Personality Science*, 194855062211441. https://doi.org/10.1177/19485506221144148

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* [dataset]. 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

Norman, J. B., Franco, M. G., & Chen, J. M. (2023). Multiracial individuals' experiences of rejection and acceptance from different racial groups and implications for life satisfaction. *The Journal of Social Psychology*, *163*(4), 459–479. https://doi.org/10.1080/00224545.2021.1996322

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, David. 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), e12392. https://doi.org/10.1111/spc3.12392

Pauker, K., Rule, N. O., & Ambady, N. (2010). Ambiguity and social perception. In *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* (Second edition). 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 W. Damon & R. M. Lerner (Eds.), *Handbook of Child Psychology*. John Wiley & Sons, Inc. 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. (2021). Gender/sex diversity beliefs: Scale construction, validation, and links to prejudice. *Group Processes & Intergroup Relations*, 136843022098759. 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: Boys Will Be Boys; Cows Will Be Cows. *Child Development*, *80*(2), 461–481. https://doi.org/10.1111/j.1467-8624.2009.01272.x

Vries, K. M. (2012). Intersectional Identities and Conceptions of the Self: The Experience of

Transgender People: Intersectional Identities. *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]. Williams Institute, UCLA School of Law. 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. (n.d.). *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

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

Table of Contents

**Additional Information about Studies 1a and 1b.....102**
Study 1a and 1b Detailed Method.....102
Study 1a and 1b Stimuli.....102
Study 1a Analyses.....103
Study 1b Analyses.....105
Studies 1a & 1b Brief Summary of Findings.....106
**Additional Information about Study 2.....107**
Study 2 Detailed Method.....107
Study 2 Stimuli.....107
Study 2 Analyses.....107
Study 2 Brief Summary of Findings.....109
**Additional Information about Study 3.....109**
Study 3 Detailed Method.....109
Study 3 Pilot Study.....110
Study 3 Stimuli.....110
Study 3 Demographics Table.....110
Study 3 Analyses.....112
Study 3 Brief Summary of Findings.....124
**Additional Information about Study 4.....124**
Study 4 Detailed Method.....124
Study 4 Stimuli.....127
Study 4 Analyses.....127
Study 4 Brief Summary of Findings.....132
**Additional Information about Study 5.....132**
Study 5 Detailed Method.....132
Study 5 Stimuli.....133
Study 5 Analyses.....133
Study 5 Brief Summary of Findings.....139

## Additional Information about Studies 1a and 1b

### *Study 1a and 1b Detailed Method*

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

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

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

### *Studies 1a and 1b Stimuli*

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

pairs served as the end points, which were systematically morphed using WebMorph's "transform" function (DeBruine, 2018) at a 10% gradient to create 9 distinct morphs from each of the 36 pairs of faces (e.g., 10% of the female face and 90% of the male, etc; see Figure 1 in manuscript for an example set of face gender stimuli)[11]. At test, participants in Study 1a sorted all morphs from half the face pairs of each race (162 faces total), and morphs from the other half of the face pairs were given to a different set of participants in Study 1b (see https://osf.io/u69yj/?view_only=c9374011248b42fcb1c257242a775b23 for full sets of face stimuli for both studies). This meant that Study 1b was a replication with unique stimuli (in line for calls for stimulus sampling; Wells & Windschitl, 1999).

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

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

### *Study 1a Analyses*

#### *Study 1a pre-registered main analyses[12]*

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

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

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

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

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

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

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

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

***Study 1a pre-registered additional analyses***

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

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

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

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

***Study 1a exploratory analyses***

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

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

***Study 1b Analyses***

***Study 1b pre-registered main analyses***

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

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

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

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

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

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

***Study 1b pre-registered additional analyses***

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

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

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

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

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

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

***Study 1b exploratory analyses***

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

***Studies 1a & 1b Brief Summary of Findings***

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

## Additional Information about Study 2

### *Study 2 Detailed Method*

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

### *Study 2 Stimuli*

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

### *Study 2 Analyses*

### *Study 2 pre-registered main analyses*

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

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

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

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

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

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

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

***Study 2 pre-registered additional analyses***

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

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

***Study 2 exploratory analyses***

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

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

***Study 2 Brief Summary of Findings***

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

**Additional Information about Study 3**

***Study 3 Detailed Method***

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

***Study 3 Pilot Study***

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

***Study 3 Stimuli***

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

***Study 3 Demographics Table (including cis-queer participants)***

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

**Table 1.** Demographic information for participants in Study 3

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

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

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

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

***Study 3 Analyses***

***Study 3 pre-registered main analyses***

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

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

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

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

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

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

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

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

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

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

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

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

***Study 3 pre-registered additional analyses***

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

*p < .05.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

***Study 3 exploratory analyses***

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

### *Study 3 Brief Summary of Findings*

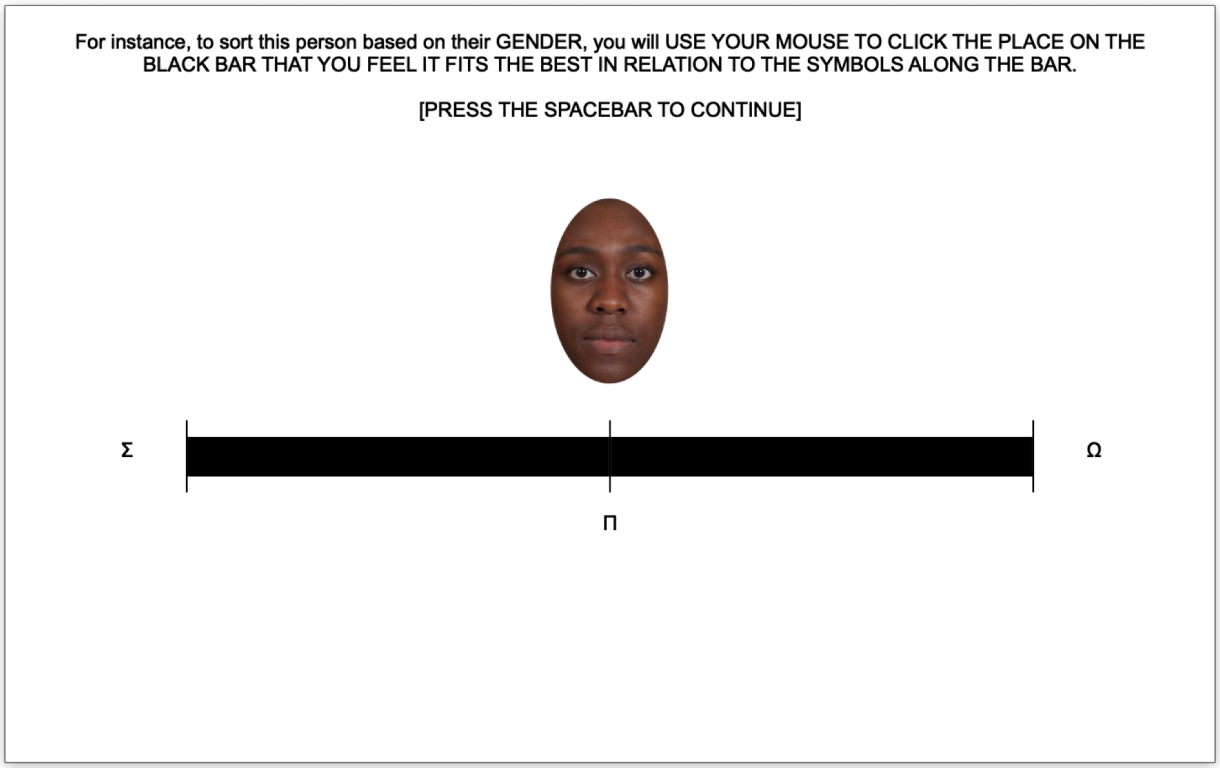
In Study 3, we asked whether people's perceptions of gender/sex differ based on experience. Specifically, we compared accuracy and latency of gender/sex judgments across gender diverse, cisgender LGBPQ+, and cisgender heterosexual participants at both "fast" and "slow" sorting conditions . For accuracy,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

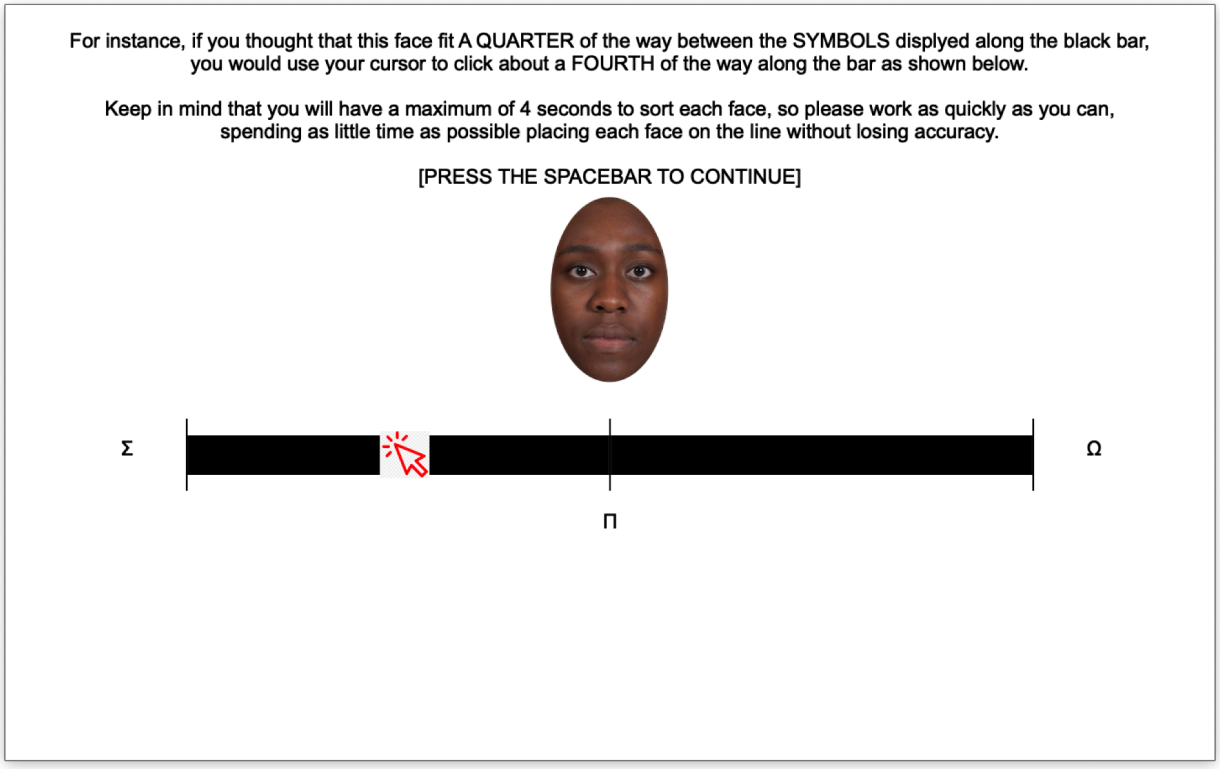
## Additional Information about Study 4

### *Study 4 Detailed Method*

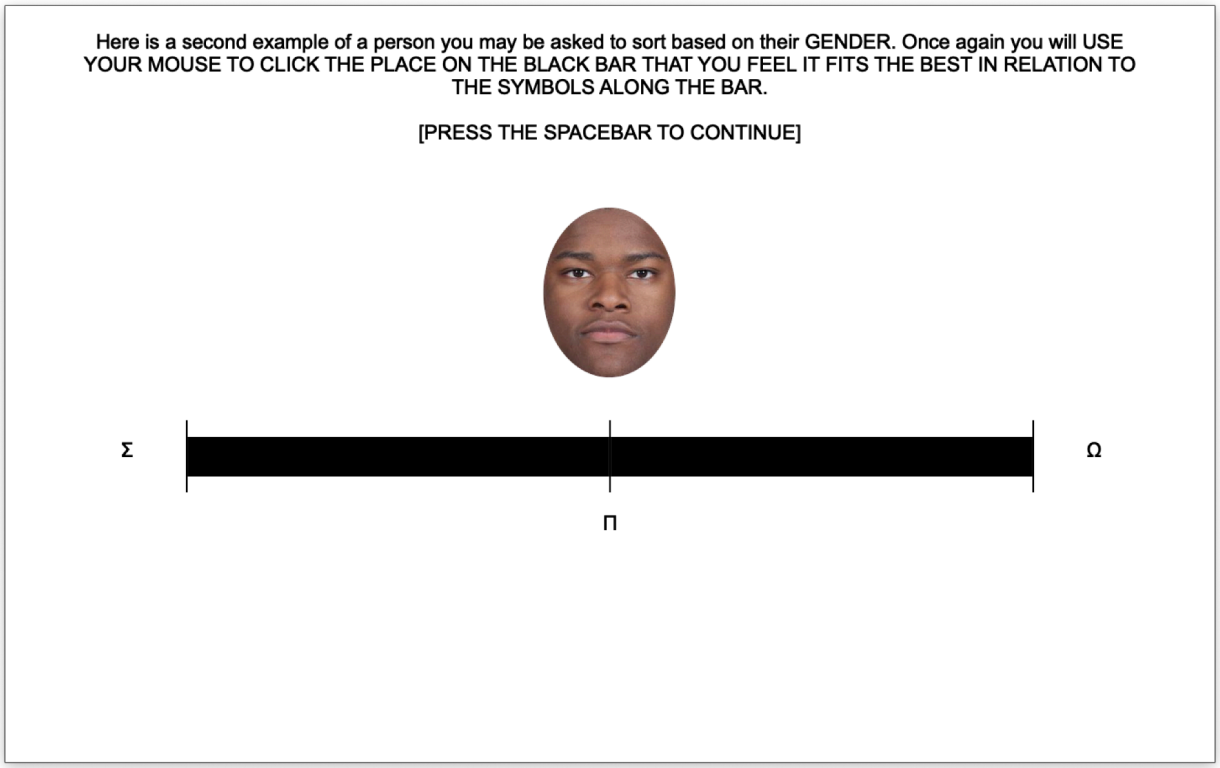
The procedure for Study 4 had two distinct differences from the set-up used in previous studies. First, participants in this study were randomly assigned to complete either a 2-category label or 3-category label version of the task. Second, instead of words (e.g. female or male), all the category labels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

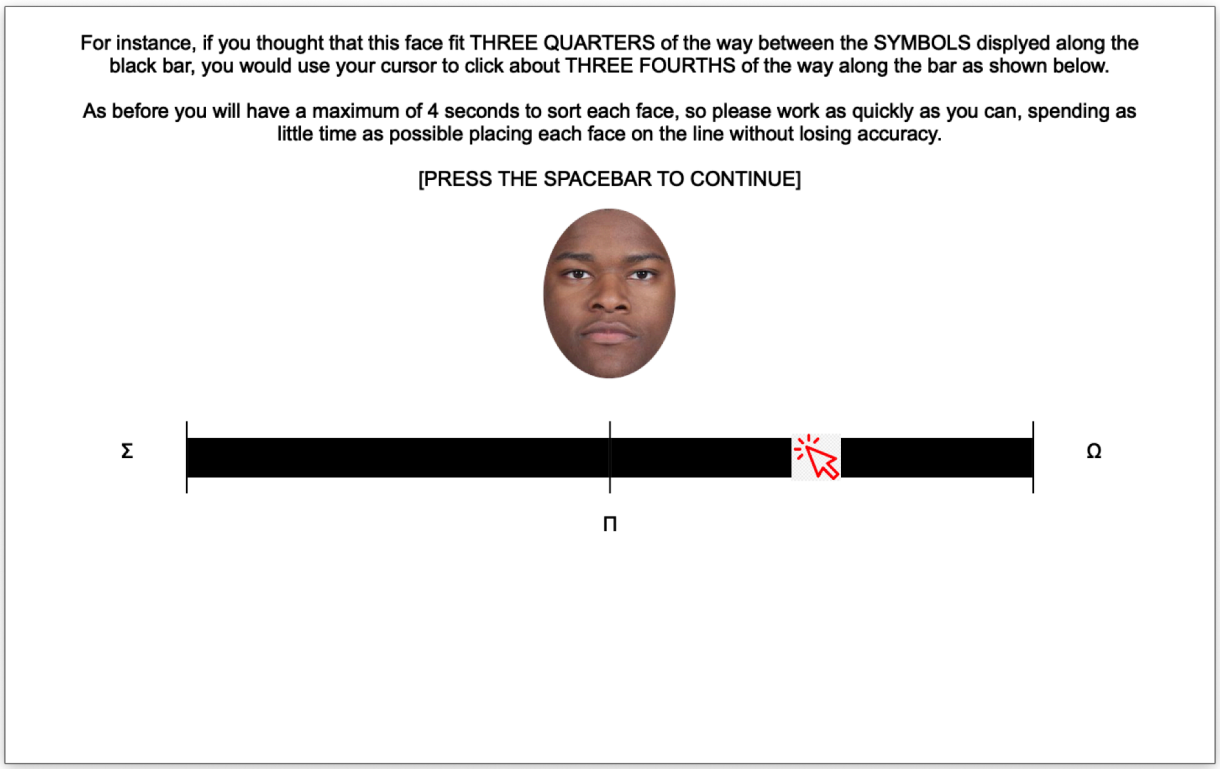
A. Guidance on placing a 25% morph exemplar.



B. Guidance on placing a 75% morph exemplar.



C. Faces and symbols mapped out on continuum

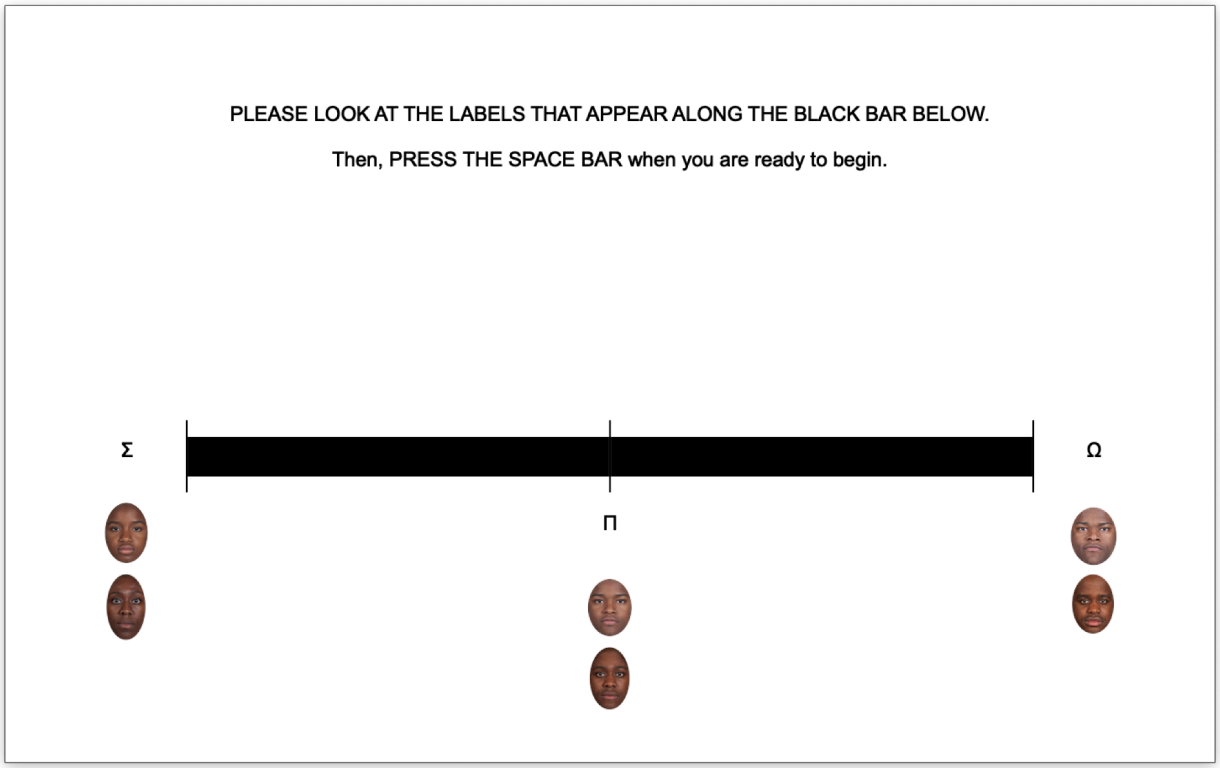



**Figure S1.** Example training received by participants in the 3-category condition of Study 4

***Study 4 Stimuli***

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

***Study 4 Analyses***

***Study 4 pre-registered main analyses***

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

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

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

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

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

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

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

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

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

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

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

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

***Study 4 pre-registered additional analyses***

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

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

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

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

***Study 4 exploratory analyses***

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

***Study 4 Brief Summary of Findings***

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

## Additional Information about Study 5

***Study 5 Detailed Method***

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

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

***Study 5 Stimuli***

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

***Study 5 Analyses***

***Study 5 pre-registered main analyses***

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

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

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

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

***Study 5 pre-registered additional analyses***

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

***Study 5 Brief Summary of Findings***

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

**Perceptions of nonbinary people as an intermediate group**

S. Atwood[1,^] & Kristina R. Olson[1,2]

[1]Department of Psychology, Princeton University, Princeton, NJ 08540 USA.

[2]Department of Psychology, University of Washington, Seattle, WA 98105 USA.

S. Atwood: https://orcid.org/0000-0003-2889-9575

Kristina R. Olson: https://orcid.org/0000-0002-1111-544X

**Author Note**: This study was funded by Princeton University and a National Science Foundation Graduate Research Fellowship awarded to S. Atwood. Subsets of these data have been presented at The Society for the Psychological Study of Social Issues and at social psychology seminars at Princeton University and the University of Delaware.

^Please direct correspondence concerning this article to: satwood@princeton.edu

**Abstract**

Existing research on conceptualization of nonbinary people is scant. Building off of previous literature on 'intermediate' groups (i.e. groups who are characterized as being 'in-between' other groups along the same identity dimension such as bisexual or biracial people; Burke et al., 2023), we first establish in a pilot study that cisgender, heterosexual people view nonbinary people as intermediate to the gender groups of men and women. Then, in Studies 1-3, we find that cisgender heterosexual people think about nonbinary individuals as a less socially legitimate group, a less stable group, and a less liked group than better known gender groups (including men and women, transgender and cisgender people, or transgender men and women) - consistent with past work suggesting that intermediate groups are often viewed in these ways. In our final study, we find that the extent to which participants think of nonbinary people less favorably than cisgender men and women negatively predicts their endorsement of nonbinary-inclusive policies. Collectively, our findings demonstrate that nonbinary individuals may not be recognized as a legitimate and stable identity group in society, and that this may hinder people's support for policies that affirm nonbinary people.

Keywords: nonbinary gender, intermediate groups, policy support

Not all social groups are seen as equally legitimate. Past work has demonstrated that some social groups that are seen as being 'in-between' other groups along the same social dimension (e.g. bisexual people relative to gay people and straight people) and that these *intermediate groups* are often perceived as less 'socially real', less stable, and less well-liked (Burke et al., 2023). The current work builds on these findings by asking whether cisgender heterosexual people perceive the category 'nonbinary' to be an intermediate group. If so, we then ask if this group is perceived as less stable, legitimate, and liked than other gender groups. Finally, we investigate the extent to which these beliefs are associated with support for nonbinary-affirming policies and whether contact with nonbinary people is associated with lower levels of these beliefs.

**Perceptions of nonbinary people**

Little research has examined perceptions of nonbinary people. What literature does exist on this topic has suggested that cisgender, heterosexual people's conceptualizations and understandings of nonbinary people are grounded in their ideas about transgender people (Schudson & Morgenroth, 2022). For instance, predictors of negative attitudes towards trans people (e.g. endorsement of gender essentialism) also predict negative attitudes towards nonbinary people (Morgenroth & Ryan, 2021; Schudson & van Anders, 2022), and conversely, positive attitudes towards transgender men and women are also associated with positive attitudes towards nonbinary people (Perez-Arche & Miller, 2021). Additional work demonstrates that, like transgender people, people have generally unfavorable opinions of nonbinary people and think of them as less attractive, lower status, and less competent than cisgender people (Mao et al., 2019; Jacques et al., 2022). Attitudes about nonbinary gender may differ by (binary) gender, with cisgender men reporting higher levels of prejudice towards nonbinary people than cisgender

women (Worthen, 2021). However, (quantitative) research about nonbinary people frequently conflates nonbinary and transgender identities, making it difficult to determine whether or how nonbinary people are thought about as distinct from more binary transgender people (Matsuno & Budge 2017, Schudson & Morgenroth, 2022).

**Common understandings of nonbinary identity as intermediate**

While perceptions of nonbinary people are not yet well-established, existing literature suggests that nonbinary people may be perceived as a group that falls intermediate to other gender groups like men and women. One reason that nonbinary people might be seen as an intermediate group is that scholarly and popular sources often define nonbinary in relation to dominant pairs of gender groups (e.g. men and women). For instance, past literature has described nonbinary as "An umbrella term for people who don't identify as exclusively male or female" (Richards et al., 2016) or "gender and/or sex identities (gender/sex) that are not exclusively female or male" (Schudson & Morgenroth, 2022). Other definitions directly include the possibility that nonbinary falls between other genders such as "a term that defines several gender identity groups…whose gender identity falls between or outside male and female identities" (Matsuno & Budge, 2017).

Paralleling the scholarly literature, the Human Rights Campaign defines nonbinary as "an adjective describing a person who does not identify exclusively as a man or a woman" and also that "non-binary people may identify as being both a man and a woman, somewhere in between, or as falling completely outside these categories" (Human Rights Campaign, n.d.). Thus, like other intermediate groups, like multiracial people, nonbinary people seem to be most commonly understood in relation to, and in-between non-intermediate groups along the same social

dimension. As a result, people who are not nonbinary may come to think of nonbinary people as a group that is intermediate to other gender groups.

**Intermediate groups**

The work of Burke et al (2023), found that in the domains of sexuality (e.g. bisexual people) and race (e.g. biracial people), intermediate groups are perceived as less socially legitimate and as having less stable identities in comparison to groups that are not considered intermediate (such as gay people, White people, etc.; Burke et al., 2023). These effects were accentuated for individuals who were highly identified with the dominant group (i.e. White identity for race and heterosexual identity for sexuality) on both dimensions. In the domain of sexuality, Burke and colleagues also found that participants evaluated bisexual individuals more negatively than gay or straight individuals, and that this effect was mediated by participants' perceptions of both social realness and stability.

There are some reasons to think that these perceptions might also apply to nonbinary people. For instance, public opinion polls have demonstrated that many Americans hold views about gender that are incompatible with the existence of nonbinary gender. For instance, a poll by Pew Research found that 60% of Americans said a person's gender identity is determined by their assigned sex at birth (Parker et al., 2022), while a separate poll by the Public Religion Research Institute found that nearly two thirds (65%) of Americans believe there are only two genders (Public Religion Research Institute, 2023), thus precluding the existence of nonbinary as a valid gender identity.

Additionally, qualitative research indicates that people may think of nonbinary as a fake identity that exists as a result of confusion about gender (Pulice-Farrow et al., 2020), suggesting that in addition to not being real, it is also an unstable identity adopted as a phase due to

'confusion' about one's 'true' identity[13] (Mohr & Rochlen, 1999; Diamond, 2006; Worthen, 2013; Burke & LaFrance, 2016). Finally, past research has suggested that nonbinary people might be liked less than non-intermediate gender groups (Herek, 2002; Sanchez & Bonam, 2009; Tate & Youssef, 2014). Specifically, evidence suggests that cisgender people may think of nonbinary people similar to how they think about transgender people (Perez-Arche & Miller, 2021), who are known for being a negatively valued social group who are liked less than cisgender men and women (Norton & Herek, 2013; Broussard & Warner, 2018; Hoffarth & Hodson, 2018). In sum, the little that we do know about how cisgender heterosexual people think about nonbinary people suggests they might be perceived as an intermediate group and that they might, as in Burke's work with other intermediate groups, be perceived as less socially real, stable, and liked relative to other gender groups.

**Relationship between beliefs about nonbinary people and policy support**

Beliefs about social groups are impactful because they are often (but not always) aligned with opinions on policies that affect those groups (Pratto et al., 1998; Swim et al., 1995 cf, see Krysan, 2000; Tuch & Hughes, 2011; Taylor & Parcel, 2019). For instance, previous research has demonstrated that people who are higher in hostile sexism (and thus have strong negative feelings towards women) are less likely to be supportive of women's rights (Masser & Abrams, 1999; Campbell et al., 1997). Complementarily, individuals who are supportive of transgender rights cite their attitudes about trans people as a key reason for their support for trans-inclusive policies (Morgenroth, Axt, & Westgate, 2024), while individuals higher in transphobia are less

[13] However, this idea stands in contrast to nonbinary people's understandings of their own gender identities which they often characterize as feeling 'unintelligible' within the context of the gender/sex binary, and may use as a tool to build community and/or resilience in the face of oppression (Schudson et al., 2019; Barsigian et al., 2020; Osborn, 2022)

like to support transgender rights (Miller, Flores, et al., 2020; Flores, Haider-Markel, et al., 2018).

In the current work, we examine the relation between beliefs about nonbinary people and preferences about policies that impact nonbinary people. Specifically, we ask whether cisgender heterosexual participants' perceptions of nonbinary people as less stable, legitimate and/or liked are related to their likelihood of supporting policies that affirm nonbinary people's basic rights (e.g., support for allowing nonbinary people to have gender-neutral markers on official identification documents, requiring places of public accommodation to have gender-neutral restrooms, requiring public health coverage for gender-affirming healthcare for nonbinary people, etc.). Here, previous research has found that thinking about LGBTQ+ identities as more distinct from other groups was related to greater support for policies that benefit that group (e.g. government-funded LGBTQ+ counseling services; Roberts et al., 2017). Thus, we would expect that perceiving nonbinary people as more socially real and distinct from other gender identities may also be associated with more support for nonbinary-affirming policies.

**Role of contact with nonbinary people**

Finally, we consider the role of contact with nonbinary people in shaping binary people's ideas about nonbinary people as well as their support for nonbinary affirming policies. Past work has demonstrated that greater contact with transgender people is associated with more positive attitudes and greater acceptance of transgender people (e.g. less anti-trans prejudice; Perez-Arche & Miller, 2021). Similarly, cisgender people who know transgender people tend to support trans rights more and tend to be against anti-trans laws (Barbir et al., 2017; Reimer et al., 2017; Hässler et al., 2020). Taken together, it's plausible that contact with nonbinary people may also be related to attitudes around nonbinary people and policies that impact them. If true, participants

who interact with nonbinary people more in their day to day lives will be more likely to think they are legitimate gender group, a more stable group, and like them more. Further, those with more contact would likely support non-binary affirming policies more.

**Current research**

In a series of one pilot study and 4 larger online studies, we examine how cisgender, heterosexual people think about nonbinary people and investigate how these understandings relate to support for policies impacting nonbinary people's rights and well-being. First we examine whether nonbinary people are seen as an intermediate social group (pilot study). Then we ask if, as Burke et al (2023) demonstrated about other intermediate groups, nonbinary people are perceived as having less stable identities, being less socially legitimate, and/or are less well liked than a range of comparison groups including cisgender men/cisgender women (Studies 1a and 2), transgender men/transgender women (Study 1b), and cisgender people/transgender people (Study 1c). Then, in Study 2, we investigate whether viewing nonbinary gender as less stable, legitimate, or less well-liked is related to support for non-binary affirming policies. In Study 2 we also investigate whether contact with nonbinary people is associated with the perceived legitimacy of nonbinary people and support for nonbinary-affirming policies amongst cisgender heterosexual participants.

## Pilot Study

In order to establish whether nonbinary people were seen as an intermediate group, we ran a pilot study ($n$ = 66). Participants were asked to rate their agreement (on a scale of 1 [strongly disagree] to 5 [strongly agree]) with a set of statements. Each statement asked about the extent to which a given group falls between two other groups. These statements included but were not limited to "the extent to which the group "Nonbinary" falls between the groups

"Women" and "Men" and "the extent to which the group "Biracial" falls between the groups "White" and "Black". Overall, participants agreed that nonbinary people are between men and women ($M$ = 4.06, $SD$ = 1.08). This rating is significantly higher than the midpoint of the scale (3 [neither agree nor disagree]; $t(65) = 7.98$, $p < .001$). This rating was similar to the amount they agreed that Biracial is between Black and White ($M$ = 4.14, $SD$ = 0.88; See SI for full list of statements used). Before completing the rating tasks, participants also completed an open-ended question about what label participants would give to a group that fell between men and women. Nonbinary was the most common answer, provided by a plurality (32%) of the sample.

**Studies 1a-1c**

Having established that cisgender, heterosexual people perceive nonbinary people as an intermediate gender group, in three studies we ask whether nonbinary people are perceived as lacking legitimacy, stability, and likeability, as Burke et al. (2023) have shown to be true of other types of intermediate groups. These three studies used nearly identical methodology with the exception of the endpoints/category labels used as comparison groups (i.e. men and women in Study 1a, transgender men and transgender women in Study 1b, and transgender and cisgender people in Study 1c). We therefore summarize their results together. All three of these studies were pre-registered (see https://osf.io/fyg3h/?view_only=f036412b0d7c429aad41caca0fc319d8 for Studies 1a[14], 1b & 1c). In the main paper, we present a subset of the preregistered analyses. However, to ensure full transparency, data and code for all preregistered analyses are available on OSF (https://osf.io/36jz4/?view_only=ff5ca4ee537546339468fc09c8d3e641) and preregistered analyses omitted from the main manuscript are reported in the Supplementary Material. All analyses were conducted using R [Version 4.3.2] software (R Core Team, 2023) via

[14] The pre-registration for Study 1a also contains preregistered analyses for an additional sample of individuals who were asked about the legitimacy, stability, and likeability of gay/lesbian, straight/heterosexual, and bisexual people (a direct replication of Burke et al., 2023). The results of this sample are included in the Supplementary Material.

RStudio [Version 2023.09.1+494]. Analyses were performed using the following packages: tidyverse (v2.0.0, Wickham, 2023), knitr (v1.45, Xie, 2023), papaja (v0.1.2, Aust & Barth, 2023), ggpubr (v0.6.0 Kassambara, 2023a), lme4 (v1.1-35.1, Bates et al., 2024), and rstatix (v0.7.2, Kassambara, 2023b).

**Methods**

*Participants*. In Study 1a, a convenience sample of 61 participants completed the study on Prolific Academic in Spring 2023. Our sample size was preregistered in advance of data collection (aiming for 59 eligible participants to give us 95% power to detect an effect of d = .49 or greater) and we excluded participants based on pre-registered exclusion criteria (see preregistration at https://osf.io/tha9f?view_only=f036412b0d7c429aad41caca0fc319d8). In total, these criteria excluded 3 individuals, which left us with 58 eligible participants in Study 1a (see Table 1 for demographic information and Table 2 for exclusion information).

In Studies 1b and 1c, a convenience sample of 76 and 75 participants, respectively completed the study on Cloud Research in Summer 2023. As before, we preregistered our exclusion criteria and sample size in advance of data collection (see https://osf.io/dnzv6?view_only=f036412b0d7c429aad41caca0fc319d8), again aiming for 59 eligible participants for each study to give us 95% power to detect an effect of d = .49 or greater (the average effect size reported Burke et al., 2023). In total, we excluded 7 individuals, which left us with 74 eligible participants in Study 1b and 70 eligible participants in Study 1c (see Table 1 for demographic information and Table 2 for exclusion information).

Table 1. Demographic information for all studies

| Study Number (total *n* per study) | Study 1a (*n* = 58) | Study 1b (*n* = 74) | Study 1c (*n* = 70) | Study 2 (*n* = 178) |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Current Gender | | | | |
| Woman | 69% | 48% | 32% | 52% |
| Man | 31% | 52% | 68% | 48% |
| Race | | | | |
| Asian | 7% | 10% | 8.5% | 11% |
| Black | 3.5% | 10% | 10% | 9.5% |
| Latinx | 3.5% | 3% | 8.5% | 9.5% |
| Multiracial | 7% | 5% | 8.5% | 8% |
| White | 79% | 70% | 59.5% | 62% |
| Other/Did Not Disclose | . | 2% | 5% | . |
| Age | | | | |
| Mean (in years) | 40.1 | 41.1 | 36.0 | 40.6 |
| SD (in years) | 14.6 | 13.9 | 11.9 | 13.5 |

Table 2. Exclusion information for all studies based on preregistered criterion

| Study Number (total exclusions per study) | Study 1a ($n$ = 3) | Study 1b ($n$ = 2) | Study 1c ($n$ = 5) | Study 2 ($n$ = 7) |
|---|---|---|---|---|
| Survey time too long (>4x mean time) | . | 1 | 2 | 1 |
| Survey time too short (<.25x mean time) | . | . | . | . |
| Failed Attention/Honesty Check | 1 | . | . | . |
| Did not identify as cisgender | . | . | . | . |
| Did not identify as heterosexual | 2 | 1 | 3 | 6 |

*Notes:* A cell with a dot denotes that no participants were excluded for a given preregistered criterion. During data collection, we set gender-based filters to recruit only cisgender people, and also included a gender-based preregistered exclusion criteria.

*Procedure.* In all studies, participants rated three gender groups on three dimensions using measures adapted directly from Burke and colleagues (2023). First, they rated each group on social realness and stability by responding to a series of 9 questions (5 questions about social realness; α's ≥ .86 and 4 questions about stability; α's ≥ .64). To capture the extent to which they believed a given gender group qualifies as a real social group, participants answered questions like "Not every collection of people is considered to be a true 'group.' Do women qualify as a group?" To capture the extent to which they believed a given social group represents a true/stable identity, participants answered questions like "Being a woman is a stable identity". For questions about social realness and stability, the order of gender groups was randomized and

within each gender group all questions were presented in a random order and on a scale from 1 [strongly disagree] to 7 [strongly agree]. Social realness and stability were calculated by taking the mean of these 5 (for social realness) and 4 (for stability) items, respectively. After answering questions about social realness and stability, participants rated how much they liked each group by responding to a thermometer question ranking each group from 0 [very unfavorable] to 100 [very favorable]. In Study 1a, participants provided these ratings for (cisgender) men[15], (cisgender) women[3], and nonbinary people. In Study 1b participants provided ratings for transgender men, transgender women, nonbinary people and in Study 1c, for transgender people, cisgender people, and nonbinary people, respectively. After providing their ratings, participants in all three studies completed a brief attention check and demographic questionnaire. All data were collected via Qualtrics.

**Results**

Using a series of one-way ANOVAs comparing perceptions of social realness within participants across gender groups, we found significant differences in average ratings of social realness (Study 1a: $F(1.12, 64) = 27.82$, $p < .001$, generalized eta squared ($\eta^2_g$) = 0.224; Study 1b: $F(1.34, 98.16) = 26.81$, $p < .001$, $\eta^2_g = 0.052$; and Study 1c: $F(1.52, 105) = 11.89$, $p < .001$, $\eta^2_g = 0.060$). Follow-up paired t-tests with Holm-Bonferroni corrections showed that nonbinary people were rated as less socially real than men ($p < .001$), women ($p < .001$), transgender men ($p < .001$), transgender women ($p < .001$), cisgender people ($p = .001$), and transgender people ($p < .001$; see Figure 1)[16]. A second series of one-way ANOVAs found significant differences in average ratings of stability within participants across gender groups (Study 1a: $F(1.17, 66.74) =$

[15] Note that participants in this Study 1a were asked about "men", "women", and "nonbinary people". Based on past work (Gallagher & Bodenhausen, 2021), we suspect the majority of participants default to imagining cisgender men and women.

[16] Here and throughout we focus on comparisons that include nonbinary people; all non-focal comparisons are included in the supplementary material.

54.98, $p < .001$, $\eta^2_g = 0.424$; Study 1b: $F(1.26, 92.15) = 30.27$, $p < .001$, $\eta^2_g = 0.060$; and Study 1c: $F(1.38, 95.14) = 87.38$, $p < .001$, $\eta^2_g = 0.361$). Follow-up paired t-tests with Holm-Bonferroni corrections showed that participants rated nonbinary people's identity as less stable than men ($p < .001$), women ($p < .001$), transgender men ($p < .001$), transgender women ($p < .001$), and cisgender people's identities ($p < .001$; see Figure 2). The stability of nonbinary people's identity was not seen as significantly different from that of transgender people's identity - when the comparison to both was cisgender (Study 1c; $p > .05$, also see Figure 2, Panel C). Finally, one-way ANOVAs showed significant differences in likability within participants across gender groups (Study 1a: $F(1.4, 79.86) = 27.04$, $p < .001$, $\eta^2_g = 0.228$; Study 1b: $F(1.56, 112.21) = 8.099$, $p = .001$, $\eta^2_g = 0.01$; and Study 1c: $F(1.33, 91.80) = 45.29$, $p < .001$, $\eta^2_g = 0.206$). Follow-up paired t-tests with Holm-Bonferroni corrections showed that participants liked nonbinary people less than men ($p < .001$), women ($p < .001$), transgender men ($p = .004$), transgender women ($p = .016$), and cisgender people's ($p < .001$; see Figure 3). However, nonbinary people were not significantly less likable than transgender people when the cisgender comparison was presented in the study (Study 1c; $p > .05$, also see Figure 3, Panel C).

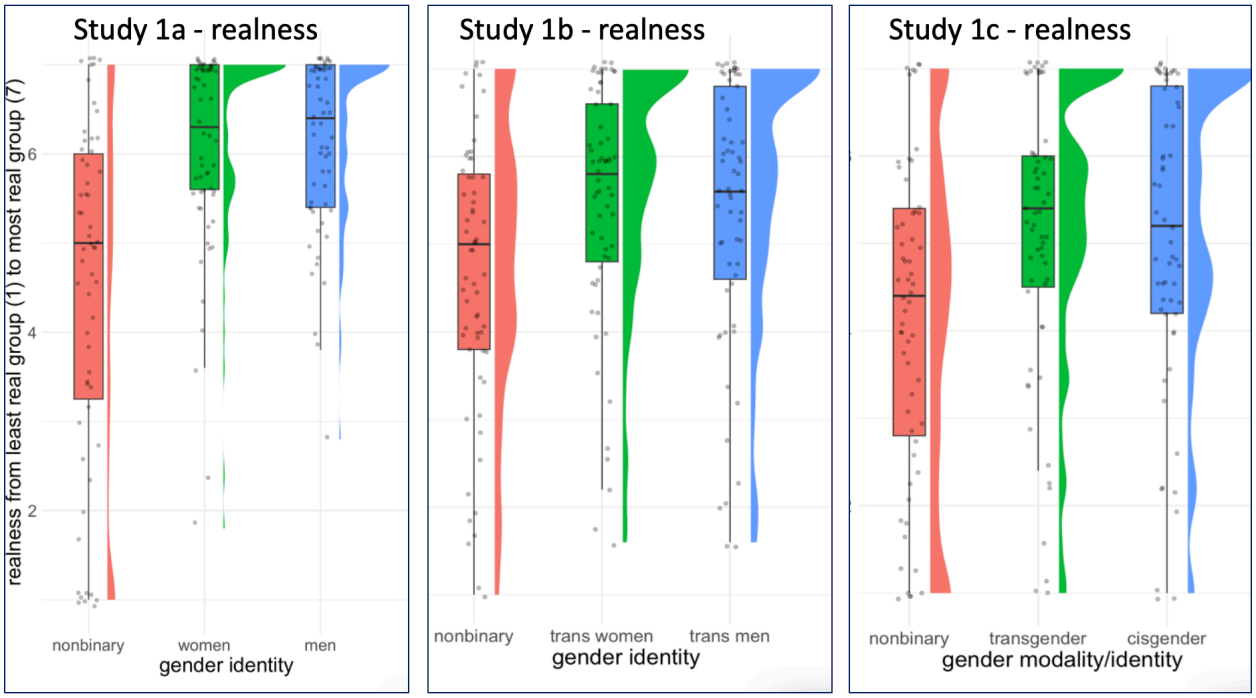



Figure 1. Participants' ratings of the social legitimacy of gender groups in studies 1a-1c.

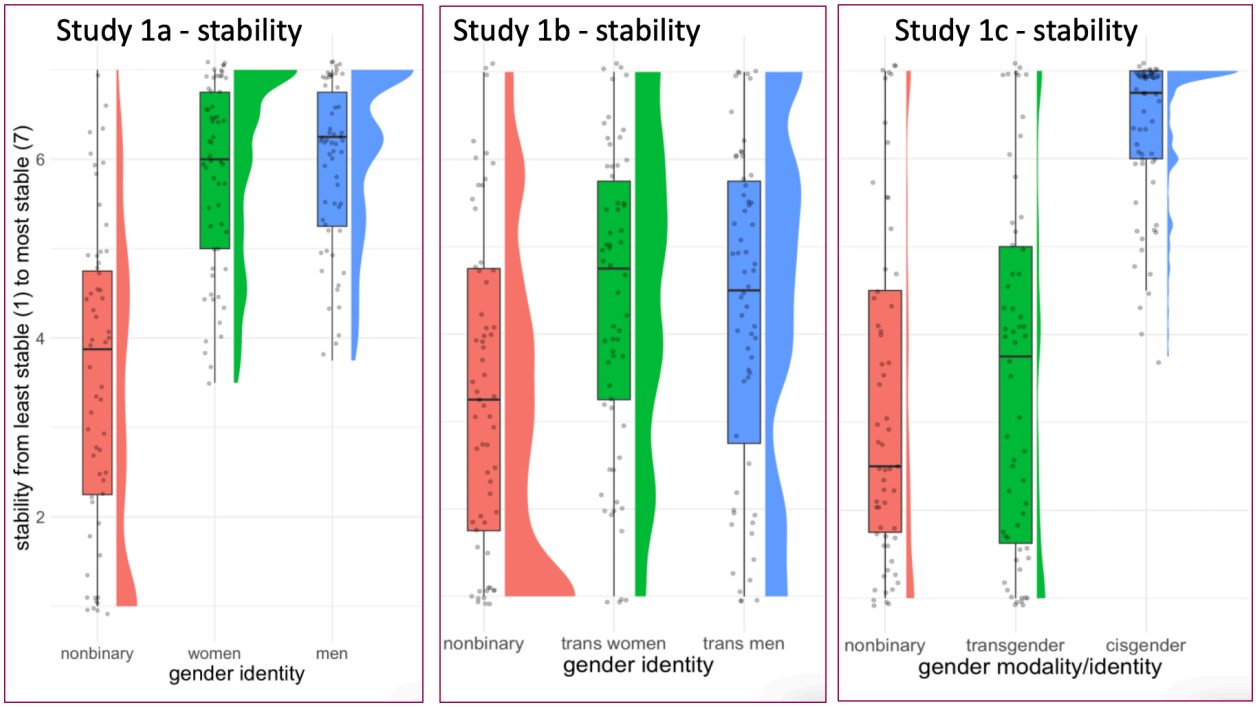



Figure 2. Participants' ratings of the stability of gender groups in studies 1a-1c.

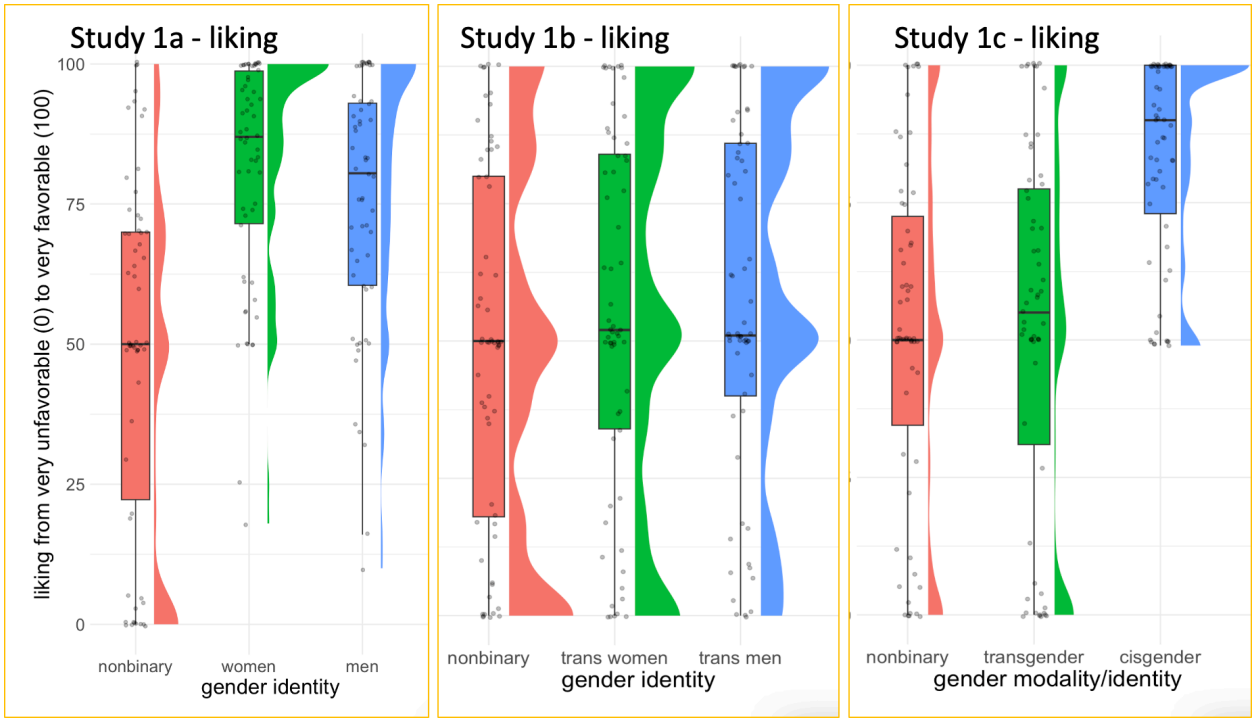



Figure 3. Participants' ratings of liking of gender groups in studies 1a-1c

**Discussion**

Across three studies, we found that nonbinary people were perceived as less socially real than men and women, transgender men and women, and cisgender and transgender people. This finding aligns with previous work (Burke et al., 2023) suggesting that other intermediate groups are also seen as less socially real. Additionally, we found that, similar to other intermediate groups, nonbinary people's identities were perceived as less stable and nonbinary people were less liked in comparison to almost all other gender categories tested. The one exception was transgender people; in this case the participants did not rate the stability or favorability of nonbinary and transgender people differently.

## Study 2

Given evidence to suggest that nonbinary people were perceived as less stable, liked, and socially real than other gender categories, our next study investigates the extent to which

people's biases against nonbinary people predict their level of support for policies that affirm nonbinary people's identities. We also investigate the role of contact with nonbinary people - specifically looking at whether or not contact with nonbinary people is related to participants' bias against nonbinary people and/or their support for nonbinary affirming policies.

**Methods**

*Participants*. In Study 2, a convenience sample of 185 participants completed the study on Prolific Academic in Winter 2024. Again, we preregistered our exclusion criteria and sample size in advance of data collection, aiming for 178 eligible participants to give us 90% power to detect correlations of r = .24 (see: https://osf.io/wy7kz?view_only=f036412b0d7c429aad41caca0fc319d8). In total, we excluded 7 individuals, which left us with 178 eligible participants in Study 2 (see Table 1 for demographic information and Table 2 for exclusion information).

*Procedure.* In this study, participants rated men, women, and nonbinary people on the same three dimensions as in Studies 1a-1c (social realness, stability, and liking). Participants also completed a measure of support for non-binary affirming policies (on a scale of 1 [strongly disagree] to 7 [strongly agree]) including support for gender-neutral markers on official documents, requiring gender neutral restrooms in places of public accommodation, public health coverage for gender-affirming healthcare, etc. (overall $\alpha = .85$; for full list of items see Table 3).

While there were no conditions in this study, we randomized the order in which participants completed these measures such that half of participants responded to questions about perceptions of different gender groups followed by questions about gender-affirming policies for nonbinary people, while the other half of participants did the reverse. Following this, participants completed a measure of contact with nonbinary individuals adapted from Hoffarth & Hodson

(2018), where they reported how much contact they had with nonbinary people (on a scale from 1 [never] to 7 [very frequently]), how often they had contact with nonbinary people (on a scale from 1 [no contact] to 7 [a great deal of contact]), average hours per week spent with nonbinary people, and the number of nonbinary people they know (for the purposes of standardization across different scales, we calculated z-scores for each item and then took the average for each participant; overall α across z-scored items = .89). Finally, participants completed a brief attention check and demographic questionnaire. All data were collected via Qualtrics.

Table 3. Items used to measure of support for nonbinary affirming policies

| |
|---|
| - Official documents such as driver's licenses and passports should allow a gender neutral X option, in addition to M and F options. |
| - All babies should be initially assigned a gender neutral marker X on their birth certificate that they can change later in their lives based on their identity. |
| - Places of public accommodation (e.g. retail stores, recreational facilities, educational institutions, etc.) should be required to have gender neutral restrooms |
| - Health insurance should cover gender affirming services needed by non-binary people (e.g. hormones, surgery, etc.). |
| - Gender identity should be a protected class meaning that the law should protect people of all genders from discrimination in employment, housing, credit, education, public spaces and services, federally funded programs, and jury service. |

*Note.* Items were presented in a random order and participants responded on a scale of 1 (strongly disagree) to 7 (strongly agree) to indicate their degree of agreement with the policy.

**Results**

Given that we did not find notable differences in the perceptions of men and women in prior studies, in this study we preregistered to combine those means and conduct comparisons of nonbinary vs. binary (men + women) identities. Conceptually replicating findings from Study 1a, a series of paired t-tests confirmed that participants thought of nonbinary people as less socially real ($t(177) = 11.59$, $p < .001$), less stable ($t(177) = 16.16$, $p < .001$), and less likable ($t(177) = 11.19$, $p < .001$) than binary gender groups (defined as the average response across men and women; see Figure 4).

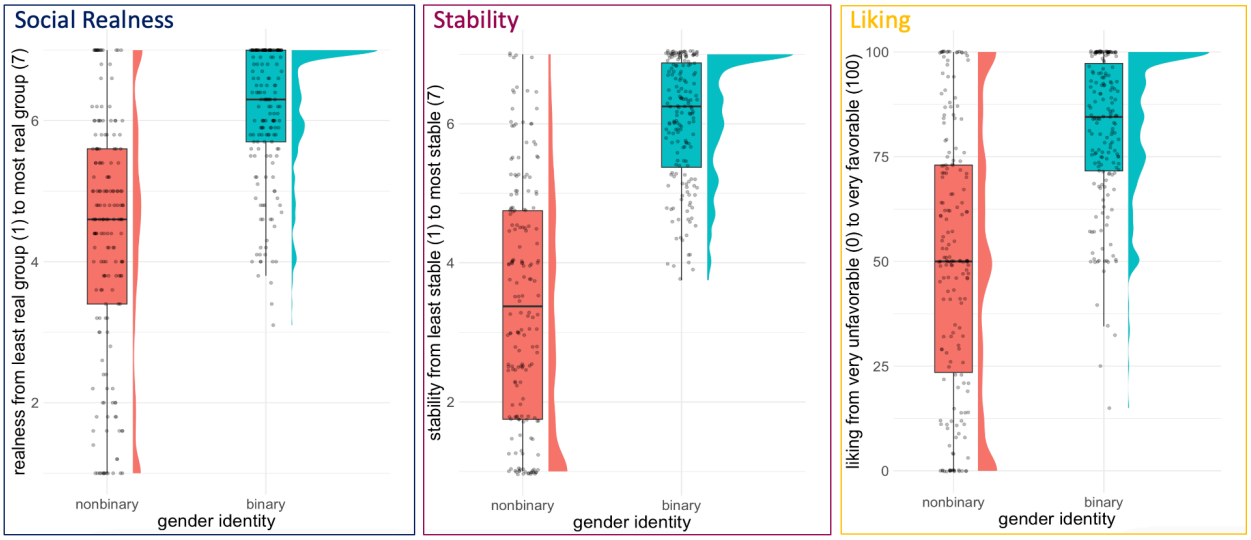



Figure 4. Participants' ratings of social realness, stability, and liking of nonbinary vs. binary gender groups.

To investigate the relation between participants' biases against nonbinary people and levels of support for policies that affirm nonbinary people's identities, we ran a series of three linear regressions predicting policy support from bias against nonbinary people in the domains of social realness, stability, and liking. In all three cases, support for nonbinary affirming policies was negatively associated with bias against nonbinary people. Specifically, participants who reported less bias against nonbinary people in the domains of social realness ($b$ = -0.55, 95% CI [−0.63,−0.46], $t(176)$ = -12.10, $p < .001$), stability ($b$ = -0.57 95% CI [-0.64, -0.49], $t(176)$ = -14.71, $p < .001$), and liking ($b$ = -0.03, 95% CI [-0.04, -0.03], $t(176)$ = -15.21, $p < .001$) also

demonstrated greater support for nonbinary affirming policies (see Figure 5).

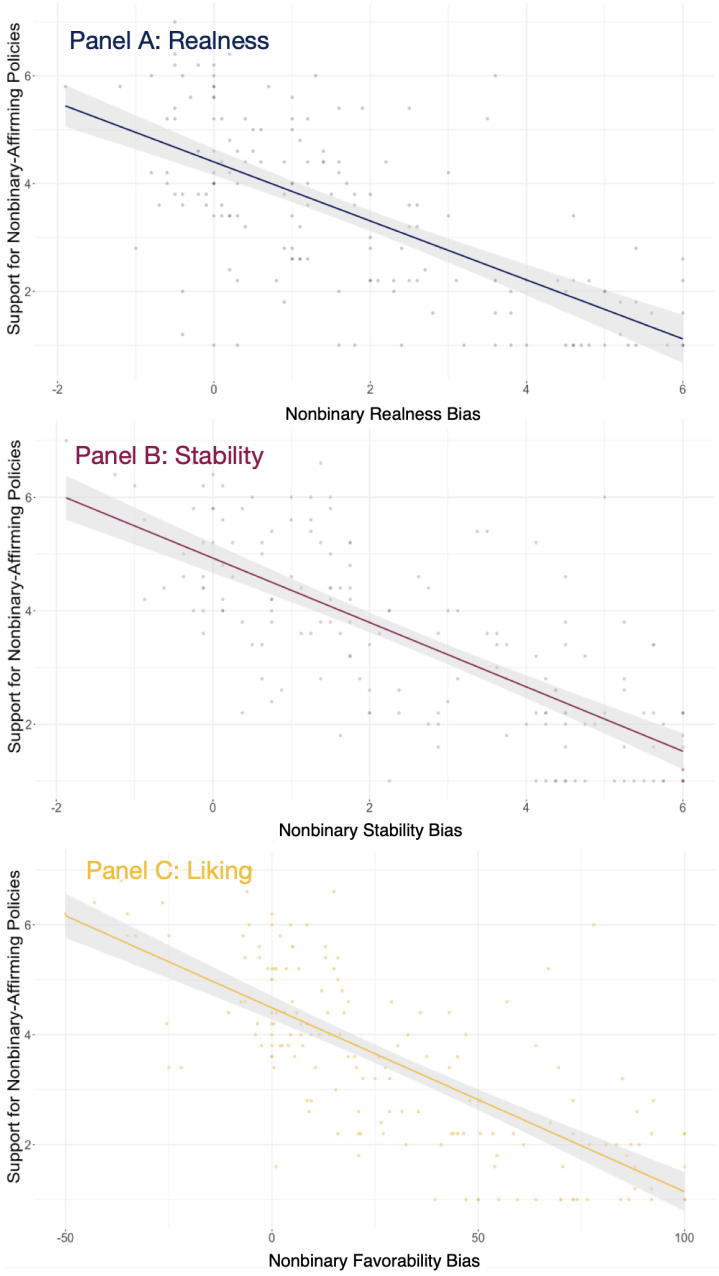



Figure 5. Relationship between support for nonbinary affirming policies and bias against nonbinary people in the domains of realness (Panel A), stability (Panel B), liking (Panel C)

Turning to the role of contact, we first investigated whether contact with nonbinary people was associated with beliefs about nonbinary people by running a series of three linear regressions predicting bias against nonbinary people in social realness, stability, and liking from

contact with nonbinary people. In each case, bias against nonbinary people was operationalized as the difference in participants' ratings between non-binary vs. binary gender on each dimension of interest (e.g. participants with a nonbinary social realness bias rated nonbinary groups as lower on social realness than binary people). We found that participants who reported greater contact with nonbinary individuals reported less bias against nonbinary (vs. binary) people in the domains of social realness ($b$ = -0.49, 95% CI [-0.83, -0.15], $t(174)$ = -2.84, $p$ = .005), stability ($b$ = -0.56, 95% CI [-0.92, -0.20], $t(174)$ = -3.07, $p$ = .002), and liking ($b$ = -10.73, 95% CI [-16.88, -4.58], $t(174)$ = -3.44, $p < .001$; also see Figure 6). Additionally, we investigated the potential relation between contact and support for nonbinary affirming policies using an additional linear regression predicting support for non-binary affirming policies from contact with nonbinary people and found that greater contact was associated with greater support for nonbinary affirming policies ($b$ = 0.32, 95% CI [0.04, 0.60], $t(174)$ = 2.29, $p$ = .023; also see Figure 6).

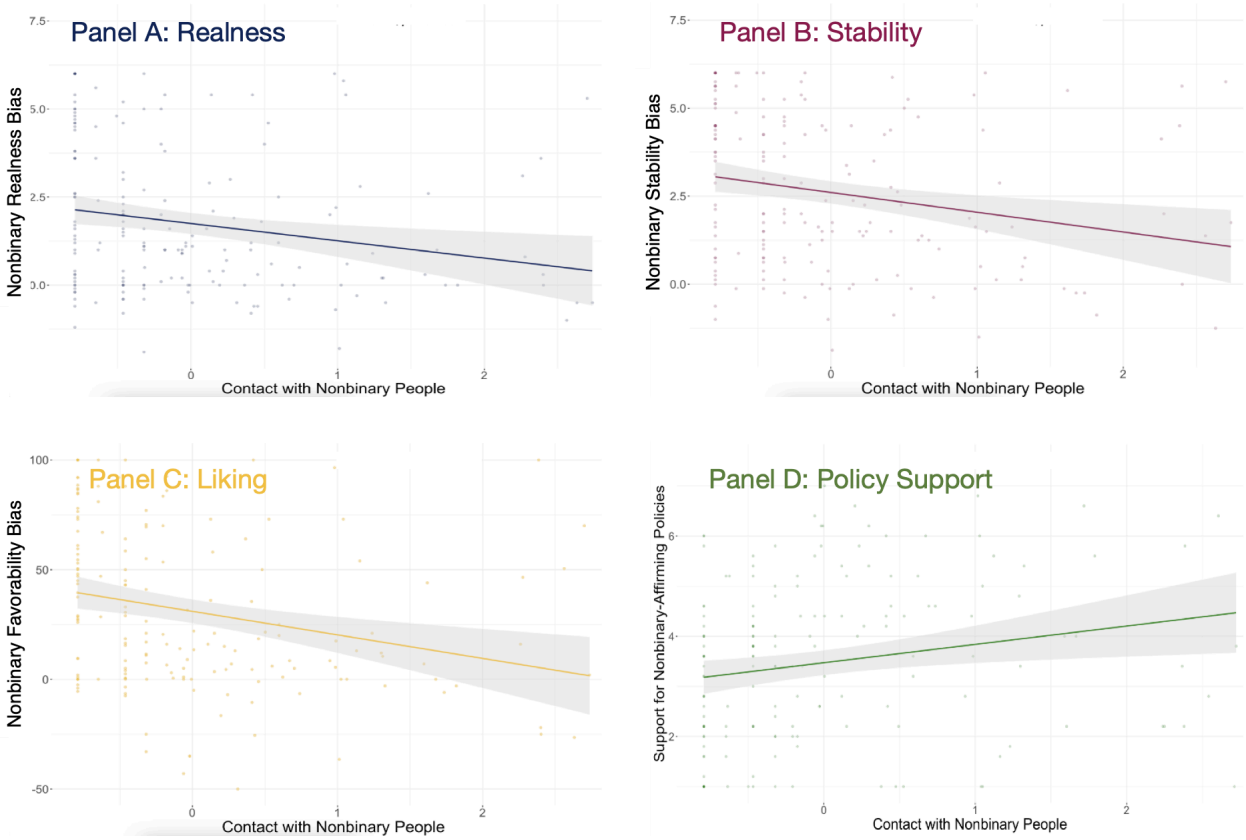



Figure 6. Relationship between contact with nonbinary people and bias against nonbinary people in the domains of realness (Panel A), stability (Panel B), liking (Panel C). Panel D depicts the relationship between contact with nonbinary people and support for nonbinary affirming policies.

**Discussion**

Study 2 expanded upon our findings from Studies 1a-1c by investigating possible correlates of thinking about nonbinary people as less real, stable, or likable than other gender groups. Specifically, we found that participants who reported greater contact with nonbinary people thought of nonbinary people as more socially real, stable, and likable (relative to cisgender men and women) than people who reported less contact with nonbinary people. Those with more contact also demonstrated greater support for nonbinary affirming policies than those with less contact. Additionally, we found that the ways participants think about nonbinary people were associated with their views on non-binary affirming policies. Specifically, the more

participants saw nonbinary people as less socially real, stable, and/or likable in comparison to cisgender men and women, the less support they demonstrated for nonbinary affirming policies.

## General Discussion

Together, these studies suggest that cisgender heterosexual people perceive nonbinary people as an intermediate group and, as they view other intermediate groups (i.e. biracial and bisexual identity; see Burke et al., 2023), they see nonbinary people as less socially real, less stable, and less likable than other gender groups. Further, these evaluations are associated with less support for policies that affirm nonbinary people.

Some of the evaluations about nonbinary people reported in these studies replicate and extend previous work on other intermediate groups. Specifically, our findings align with work suggesting that bisexual and biracial people are perceived as less socially real and less liked (Burke et al, 2023). Burke and colleagues (2023) extend an argument made by Chen and colleagues (see Chen & Hamilton, 2012; Chen et al., 2014), and posit that intermediate groups may be perceived as less real due to participant's lack of familiarity with them. Our findings from Study 2 – where participants who reported greater contact with nonbinary individuals reported less realness bias against nonbinary (vs. binary) people – align with this interpretation.

Another possibility is that cisgender participants may have perceived nonbinary people as less real because they have only recently been acknowledged in mainstream discourse, and thus may appear to be a relatively newer category in comparison to men and women. That is, for people who believe that groups are only real when they have existed throughout history, holding the (incorrect) belief that nonbinary people are a "new" group could lead them to perceive nonbinary people as less legitimate or real. Finally, intermediate groups may be perceived as less real due to lack of distinctiveness (Burke et al., 2023). Applied here, nonbinary people may be

perceived as sharing characteristics with both men and women, and so less of their own distinctive group. Finally, past work has demonstrated that regardless of its source, a lack of social realness may result in being perceived as illegitimate or inauthentic (Mohr & Rochlen, 1999), which in turn, may contribute to being less liked (Heider, 1958).

Finally, it is possible that some of the evaluations about nonbinary people could be grounded in observations or assumptions about nonbinary people's identity trajectory. For instance, one reason nonbinary identities may be perceived as less stable is because people are rarely assumed nonbinary at birth. If one assumes no one is nonbinary early in development, then anyone who is nonbinary had to experience a change or transition into this group later in life. This might lead to the perception of nonbinary people as having less stable identities. However, this logic is cisnormative and is a reflection of assuming binary gender identity as the default. Importantly, there are youth whose gender identities are not binary quite early in development (Rae et al., 2019; Gülgöz et al., 2022). We do not know of data suggesting whether these identities are any more or less stable than other gender identities over time. Further, even if nonbinary identities are less stable, this would not imply that these identities are any less legitimate or socially real than any other gender identity.

Additionally, it is important to remember that even if cisgender people think about nonbinary gender as 'in-between' the gender categories of men and women, this does not mean that this is how nonbinary people think of themselves. In fact, previous work suggests that there is no universal experience of being nonbinary (Wang et al., in prep) and that nonbinary people's understandings of their own gender are nuanced, heterogeneous, and often unintelligible within the context of the gender/sex binary (Schudson & van Anders, 2019; Barsigian et al., 2020;

Osborn, 2022). For these reasons, we believe it's particularly important for more research to focus on how nonbinary people think about their own identities.

Regardless of whether or not cisgender people's perceptions reflect how some nonbinary people think about their own identities, their perceptions are impactful and important to study because they are associated with levels of support for non-binary affirming policies. Specifically, this work suggests that thinking of nonbinary people's identities as less stable and less real, and liking nonbinary people less than cisgender men and women is associated with less support for policies that affect nonbinary people's ability to fully participate in society, such the use of an X marker on official documents, requiring gender-neutral bathrooms in places of public accommodation, public health coverage of gender-affirming healthcare for nonbinary people, etc.. This finding supports a much larger set of work suggesting that there are often, but not always, links between beliefs and policy support (Pratto et al., 1998; Swim et al., 1995; Henry & Sears, 2002; though cf. principle-policy gap, Krysan, 2000; Tuch & Hughes, 2011; Taylor & Parcel, 2019; Smith & Mayorga-Gallo, 2017).

Finally, this work contributes to literature on intergroup contact, suggesting that greater reported contact with nonbinary people is associated with at least two positive outcomes. Cisgender heterosexual participants who reported more contact with nonbinary people viewed nonbinary people as more socially real, stable, and liked, and were more supportive of nonbinary-affirming policies. This suggests that prior contact with nonbinary people may reduce these biases against nonbinary people and increase the extent to which people think of nonbinary as a legitimate gender group. However, it's also possible that individuals who already think of nonbinary people as a legitimate and stable social group, or who like nonbinary people more, are more likely to choose to interact more with nonbinary people or to simply befriend nonbinary

people more often. As with all correlational studies, we cannot determine which causal pathway, or if neither causal pathway best explains this correlation. Thus, studies that can assess causality, such as through contact-based interventions, would be helpful to establish directionality. Finally, future work might investigate potential third variables that predict both beliefs and contact or policy beliefs including (but not limited to) gender essentialism, beliefs about gender and/or sex diversity, familiarity with gender minorities, and/or ingroup bias favoring one's own gender group.

Limitations and constraints on generality

There are several constraints on the generalizability of our findings. First, due to reliance on convenience sampling, our findings are primarily representative of (mostly) White cisgender, heterosexual individuals who do studies on Prolific Academic. While limited in scope, the perceptions of White cisgender heterosexual men and women are important to understand, as they represent the demographic makeup of those who ultimately have the most social power in the United States and who are, for example, the largest voting block. As a result, their views are relevant to policies that could impact nonbinary people's everyday lives.

Relatedly, this work is focused exclusively on perceptions of nonbinary people's identity, and thus our findings do not generalize to to other gender groups and future work should address the extent to which other gender groups that could perceived as intermediate (e.g. agender, bigender, two-spirit, etc.). Future studies should assess the degree to which these findings apply across individuals with different sociocultural understandings of gender (e.g. across countries, languages, education levels, age groups, etc).

Ultimately, we hope that this work inspires further research on the perceptions of nonbinary individuals and encourages a broader exploration of gender diversity in society.

## References


Aust, F., & Barth, M. (2023). *papaja: Prepare American Psychological Association Journal Articles with R Markdown* (Version 0.1.2) [Dataset]. https://doi.org/10.32614/CRAN.package.papaja

Barbir, L. A., Vandevender, A. W., & Cohn, T. J. (2017). Friendship, attitudes, and behavioral intentions of cisgender heterosexuals toward transgender individuals. *Journal of Gay & Lesbian Mental Health*, *21*(2), 154–170. https://doi.org/10.1080/19359705.2016.1273157

Barsigian, L. L., Hammack, P. L., Morrow, Q. J., Wilson, B. D. M., & Russell, S. T. (2020). Narratives of gender, sexuality, and community in three generations of genderqueer sexual minorities. *Psychology of Sexual Orientation and Gender Diversity*, *7*(3), 276–292. https://doi.org/10.1037/sgd0000384

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2024). *lme4: Linear Mixed-Effects Models using "Eigen" and S4* (Version 1.1-35.1) [Dataset]. https://doi.org/10.32614/CRAN.package.lme4

Broussard, K. A., & Warner, R. H. (2019). Gender Nonconformity Is Perceived Differently for Cisgender and Transgender Targets. *Sex Roles*, *80*(7–8), 409–428. https://doi.org/10.1007/s11199-018-0947-z

Burke, S. E., & LaFrance, M. (2016). Lay Conceptions of Sexual Minority Groups. *Archives of Sexual Behavior*, *45*(3), 635–650. https://doi.org/10.1007/s10508-015-0655-5

Burke, S. E., Perry, S. P., Dovidio, J. F., & LaFrance, M. (2023). Distinctive negative reactions to intermediate social groups. *Journal of Applied Social Psychology*, *53*(7), 600–630. https://doi.org/10.1111/jasp.12942

Campbell, B., Schellenberg, E. G., & Senn, C. Y. (1997). Evaluating Measures of Contemporary Sexism. *Psychology of Women Quarterly*, *21*(1), 89–102. https://doi.org/10.1111/j.1471-6402.1997.tb00102.x

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., 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

Diamond, L. M. (2006). What We Got Wrong About Sexual Identity Development: Unexpected Findings From a Longitudinal Study of Young Women. In A. M. Omoto & H. S. Kurtzman (Eds.), *Sexual orientation and mental health: Examining identity and development in lesbian, gay, and bisexual people.* (pp. 73–94). American Psychological Association. https://doi.org/10.1037/11261-004

Flores, A. R., Haider-Markel, D. P., Lewis, D. C., Miller, P. R., Tadlock, B. L., & Taylor, J. K. (2018). Transgender prejudice reduction and opinions on transgender rights: Results from a mediation analysis on experimental data. *Research & Politics*, *5*(1). https://doi.org/10.1177/2053168018764945

Gallagher, N. M., & Bodenhausen, G. V. (2021). Gender essentialism and the mental representation of transgender women and men: A multimethod investigation of stereotype content. *Cognition*, *217*, 104887. https://doi.org/10.1016/j.cognition.2021.104887

Gülgöz, S., Edwards, D. L., & Olson, K. R. (2022). Between a boy and a girl: Measuring gender identity on a continuum. *Social Development*, *31*(3), 916–929. https://doi.org/10.1111/sode.12587

Hässler, T., Ullrich, J., Bernardino, M., Shnabel, N., Laar, C. V., Valdenegro, D., Sebben, S., Tropp, L. R., Visintin, E. P., González, R., Ditlmann, R. K., Abrams, D., Selvanathan, H. P., Branković, M., Wright, S., Von Zimmermann, J., Pasek, M., Aydin, A. L., Žeželj, I., … Ugarte, L. M. (2020). A large-scale test of the link between intergroup contact and support for social change. *Nature Human Behaviour*, *4*(4), 380–386. https://doi.org/10.1038/s41562-019-0815-z

Heider, F. (1958). *The psychology of interpersonal relations*. John Wiley & Sons Inc. https://doi.org/10.1037/10628-000

Henry, P. J., & Sears, D. O. (2002). The Symbolic Racism 2000 Scale. *Political Psychology*, *23*(2), 253–283. https://doi.org/10.1111/0162-895X.00281

Herek, G. M. (2002). Heterosexuals' attitudes toward bisexual men and women in the United States. *The Journal of Sex Research*, *39*(4), 264–274. https://doi.org/10.1080/00224490209552150

Hoffarth, M. R., & Hodson, G. (2018). When intergroup contact is uncommon and bias is strong: The case of anti-transgender bias. *Psychology & Sexuality*, *9*(3), 237–250. https://doi.org/10.1080/19419899.2018.1470107

Human Rights Campaign. (n.d.). *Glossary of Terms*. Human Rights Campaign. https://www.hrc.org/resources/glossary-of-terms

Jacques, S. A., Ross, D. E., & McCarty, M. K. (2022). Perceptions of Nonbinary Identifying Individuals: Through the Lens of Gender and Race. *Psi Chi Journal of Psychological Research*, *27*(1), 46–58. https://doi.org/10.24839/2325-7342.JN27.1.46

Kassambara, A. (2023a). *ggpubr: "ggplot2" Based Publication Ready Plots* (Version 0.6.0) [Dataset]. https://doi.org/10.32614/CRAN.package.ggpubr

Kassambara, A. (2023b). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests* (Version 0.7.2) [Dataset]. https://doi.org/10.32614/CRAN.package.rstatix

Krysan, M. (2000). Prejudice, Politics, and Public Opinion: Understanding the Sources of Racial Policy Attitudes. *Annual Review of Sociology*, *26*(1), 135–168. https://doi.org/10.1146/annurev.soc.26.1.135

Mao, J. M., Haupert, M. L., & Smith, E. R. (2019). How Gender Identity and Transgender Status Affect Perceptions of Attractiveness. *Social Psychological and Personality Science*, *10*(6), 811–822. https://doi.org/10.1177/1948550618783716

Masser, B., & Abrams, D. (1999). Contemporary Sexism: The Relationships Among Hostility, Benevolence, and Neosexism. *Psychology of Women Quarterly*, *23*(3), 503–517. https://doi.org/10.1111/j.1471-6402.1999.tb00378.x

Matsuno, E., & Budge, S. L. (2017). Non-binary/Genderqueer Identities: A Critical Review of the Literature. *Current Sexual Health Reports*, *9*(3), 116–120. https://doi.org/10.1007/s11930-017-0111-8

Miller, P. R., Flores, A. R., Haider-Markel, D. P., Lewis, D. C., Tadlock, B., & Taylor, J. K. (2020). The Politics of Being "Cait": Caitlyn Jenner, Transphobia, and Parasocial Contact Effects on Transgender-Related Political Attitudes. *American Politics Research*, *48*(5), 622–634. https://doi.org/10.1177/1532673X20906460

Mohr, J. J., & Rochlen, A. B. (1999). Measuring attitudes regarding bisexuality in lesbian, gay male, and heterosexual populations. *Journal of Counseling Psychology*, *46*(3), 353–369. https://doi.org/10.1037/0022-0167.46.3.353

Morgenroth, T., Axt, J. R., & Westgate, E. C. (2024). What Underlies the Opposition to Trans-Inclusive Policies? The Role of Concerns About Male Violence Versus Attitudes Toward Trans People. *Personality and Social Psychology Bulletin*, *50*(4), 533–549. https://doi.org/10.1177/01461672221137201

Morgenroth, T., & Ryan, M. K. (2021). The Effects of Gender Trouble: An Integrative Theoretical Framework of the Perpetuation and Disruption of the Gender/Sex Binary. *Perspectives on Psychological Science*, *16*(6), 1113–1142. https://doi.org/10.1177/1745691620902442

Norton, A. T., & Herek, G. M. (2013). Heterosexuals' Attitudes Toward Transgender People: Findings from a National Probability Sample of U.S. Adults. *Sex Roles*, *68*(11–12), 738–753. https://doi.org/10.1007/s11199-011-0110-6

Osborn, M. (2022). "Nobody Ever Correctly Recognizes Me": Nonbinary Presentation, Visibility, and Safety Across Contexts. In M. T. Segal & V. Demos (Eds.), *Advances in Gender Research* (pp. 51–69). Emerald Publishing Limited. https://doi.org/10.1108/S1529-212620220000033011

Parker, K., Horowitz, J. M., & Brown, A. (2022). *Americans' Complex Views on Gender Identity and Transgender Issues*. Pew Research Center. https://www.pewresearch.org/social-trends/2022/06/28/americans-complex-views-on-gender-identity-and-transgender-issues/

Perez-Arche, H., & Miller, D. J. (2021). What Predicts Attitudes toward Transgender and Nonbinary People? An Exploration of Gender, Authoritarianism, Social Dominance, and Gender Ideology. *Sex Roles*, *85*(3–4), 172–189. https://doi.org/10.1007/s11199-020-01212-3

Pratto, F., Stallworth, L. M., & Conway‐Lanz, S. (1998). Social Dominance Orientation and the Ideological Legitimization of Social Policy [1]. *Journal of Applied Social Psychology*, *28*(20), 1853–1875. https://doi.org/10.1111/j.1559-1816.1998.tb01349.x

Public Religion Research Institute. (2023). *The Politics of Gender, Pronouns, and Public Education*. Public Religion Research Institute. https://www.prri.org/research/the-politics-of-gender-pronouns-and-public-education/

Pulice-Farrow, L., McNary, S. B., & Galupo, M. P. (2020). "Bigender is just a Tumblr thing": Microaggressions in the romantic relationships of gender non-conforming and agender transgender individuals. *Sexual and Relationship Therapy*, *35*(3), 362–381. https://doi.org/10.1080/14681994.2018.1533245

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rae, J. R., Gülgöz, S., Durwood, L., DeMeules, M., Lowe, R., Lindquist, G., & Olson, K. R. (2019). Predicting Early-Childhood Gender Transitions. *Psychological Science*, *30*(5), 669–681. https://doi.org/10.1177/0956797619830649

Reimer, N. K., Becker, J. C., Benz, A., Christ, O., Dhont, K., Klocke, U., Neji, S., Rychlowska, M., Schmid, K., & Hewstone, M. (2017). Intergroup Contact and Social Change: Implications of Negative and Positive Contact for Collective Action in Advantaged and Disadvantaged Groups. *Personality and Social Psychology Bulletin*, *43*(1), 121–136. https://doi.org/10.1177/0146167216676478

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

Roberts, S. O., Ho, A. K., Rhodes, M., & Gelman, S. A. (2017). Making Boundaries Great Again: Essentialism and Support for Boundary-Enhancing Initiatives. *Personality and Social Psychology Bulletin*, *43*(12), 1643–1658. https://doi.org/10.1177/0146167217724801

RStudio Team. (2023). *RStudio: Integrated development environment for R* (Version 2023.09.1+494) [Computer software]. Posit Software, PBC. https://posit.co/download/rstudio/

Sanchez, D. T., & Bonam, C. M. (2009). To Disclose or Not to Disclose Biracial Identity: The Effect of Biracial Disclosure on Perceiver Evaluations and Target Responses. *Journal of Social Issues*, *65*(1), 129–149. https://doi.org/10.1111/j.1540-4560.2008.01591.x

Schudson, Z. C., Beischel, W. J., & Van Anders, S. M. (2019). Individual variation in gender/sex category definitions. *Psychology of Sexual Orientation and Gender Diversity*, *6*(4), 448–460. https://doi.org/10.1037/sgd0000346

Schudson, Z. C., & Morgenroth, T. (2022). Non-binary gender/sex identities. *Current Opinion in Psychology*, *48*, 101499. https://doi.org/10.1016/j.copsyc.2022.101499

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

Smith, C. W., & Mayorga-Gallo, S. (2017). The New Principle-policy Gap: How Diversity Ideology Subverts Diversity Initiatives. *Sociological Perspectives*, *60*(5), 889–911. https://doi.org/10.1177/0731121417719693

Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A. (1995). Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology*, *68*(2), 199–214. https://doi.org/10.1037/0022-3514.68.2.199

Tate, C. C., Youssef, C. P., & Bettergarcia, J. N. (2014). Integrating the Study of Transgender Spectrum and Cisgender Experiences of Self-Categorization from a Personality Perspective. *Review of General Psychology*, *18*(4), 302–312. https://doi.org/10.1037/gpr0000019

Taylor, A. J., & Parcel, T. L. (2019). Proximity and the principle-policy gap in white racial attitudes: Insight from views of student assignment policies in Wake County, North Carolina. *Social Science Research*, *78*, 95–103. https://doi.org/10.1016/j.ssresearch.2018.12.014

Tuch, S. A., & Hughes, M. (2011). Whites' Racial Policy Attitudes in the Twenty-First Century: The Continuing Significance of Racial Resentment. *The ANNALS of the American Academy of Political and Social Science*, *634*(1), 134–152. https://doi.org/10.1177/0002716210390288

Wickham, H. (2023). *tidyverse: Easily Install and Load the "Tidyverse"* (Version 2.0.0) [Dataset]. https://doi.org/10.32614/CRAN.package.tidyverse

Worthen, M. G. F. (2013). An Argument for Separate Analyses of Attitudes Toward Lesbian, Gay, Bisexual Men, Bisexual Women, MtF and FtM Transgender Individuals. *Sex Roles*, *68*(11–12), 703–723. https://doi.org/10.1007/s11199-012-0155-1

Worthen, M. G. F. (2021). Why Can't You Just Pick One? The Stigmatization of Non-binary/Genderqueer People by Cis and Trans Men and Women: An Empirical Test of Norm-Centered Stigma Theory. *Sex Roles*, *85*(5–6), 343–356. https://doi.org/10.1007/s11199-020-01216-z

Xie, Y. (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R* (Version 1.45) [Dataset]. https://doi.org/10.32614/CRAN.package.knitr

**Supplementary Material for *Perceptions of nonbinary people as an intermediate group***

Table of Contents

**Pilot Study** ..... **176**
Pilot Study Procedure ..... 176
Pilot Study Results ..... 176
**Additional Information about Study 1a** ..... **176**
Detailed Procedure ..... 176
Preregistered Analyses ..... 177
**Additional Information about Studies 1b & 1c** ..... **179**
Other Preregistered Analyses ..... 179
**Additional Information about Study 2** ..... **182**
Other Preregistered Analyses ..... 182

## Pilot Study

### *Pilot Study Procedure*

We conducted a small pilot study ($n$ = 66) in order to determine whether or not people think about nonbinary as being intermediate. For the purposes of this pilot study, we operationalized intermediate as being “between men and women”, and we tested whether nonbinary people were perceived as being intermediate in two ways. First, we asked several open-ended questions where participants were presented with a pair of groups and were asked to write in the label for a group that falls between the other two groups. In order to reduce demand effects, we asked about several pairs of groups presented in random order including Black and White, Old and Young, Rich and Poor, Man and Woman, Child and Adult, and Liberal and Conservative. After providing free-response labels, participants were asked to rate their agreement (on a scale of 1 [strongly disagree] to 5 [strongly agree]) about the extent to which they thought a given group falls between two other groups. Specifically, they were asked to rate “the extent to which the group "Biracial" falls between the groups "White" and "Black", “the extent to which the group "Middle-Aged" falls between the groups "Young" and "Old", “the extent to which the group "Middle Class" falls between the groups "Rich" and "Poor", “the extent to which the group "Nonbinary" falls between the groups "Man" and "Woman", “the extent to which the group "Teenager" falls between the groups "Child" and "Adult", and “the extent to which the group "Moderate" falls between the groups "Conservative" and "Liberal"

### *Pilot Study Results*

We evaluated whether or not people think about nonbinary as being intermediate in two ways. First, we looked at the labels generated for the group that falls between men and women calculated the percentage of people who wrote nonbinary. We found that nonbinary was the most common answer (32%), followed by transgender (23%), and hermaphrodite (9%). Other labels were generated included asexual, bisexual, child, gender fluid, gender-neutral, homosexual, intergender, intersex, man, they, undecided, and uni. Despite accounting for less than a third of the responses, participants were much more likely to spontaneously write nonbinary as a the group between men and women than they were to spontaneously write biracial (a group considered to be intermediate prior research; Burke et al., 2023), which accounted for only 9% of responses when asked for a group between Black and White. We did not examine the free-response labels generated for other pairs of groups as they did not pertain to our question about nonbinary people.

We also quantified the extent to which participants agreed (on a scale of 1 [strongly disagree] to 5 [strongly agree]) that nonbinary falls between man and woman ($M$ = 4.06, $SD$ = 1.08). A t-test revealed that participants’ ratings were significantly higher than the (neutral) midpoint of the scale (3 [neither agree nor disagree]; $t(65) = 7.98$, $p < .001$)). Additionally, the extent to which participants agreed that nonbinary fell between man and woman was similar to the amount they agreed that Biracial fell between Black and White ($M$ = 4.14, $SD$ = 0.88). We did not examine the extent to which people agreed about other potential ‘intermediate’

groups as they did not pertain to our question about nonbinary people.

## Additional Information about Study 1a

### *Detailed Procedure*

In addition to the data reported in the main paper comparing participants' perceptions of nonbinary people (as compared to men and women), Study 1a also included a second (preregistered) condition ($N = 60$)[17] aimed at directly replicating select findings from Burke, Perry, Dovidio, & LaFrance (2023)'s about participant's perceptions of the legitimacy, stability, and likability of bisexual people (as compared to gay/lesbian people and straight/heterosexual people). In this additional condition, participants rated bisexual people, gay/lesbian people, and straight/heterosexual people on the same three dimensions used in the main paper (social realness, stability, and likeability) adapted directly from Burke and colleagues (2023). Like in Studies 1a-1c in the main paper, participants rated each group on social realness and stability by responding to 9 questions presented in a random order on a scale from 1 [strongly disagree] to 7 [strongly agree]). For instance, to capture the extent to which each group qualifies as a real social group, participants answered questions like "Not every collection of people is considered to be a true 'group.' Do bisexual people qualify as a group?" and to capture the extent to which they believed a given social group represents a true/stable identity, participants answered questions like "Being bisexual is a stable identity". After answering questions about social realness and stability, participants rated how much they liked each group by responding to a thermometer question ranking each group from 0 [least favorable] to 100 [most favorable]. All preregistered analyses for this condition are included below.

***Preregistered Analyses***

We will run 3 anovas looking at overall differences in legitimacy/social realness (anova 1), stability (anova 2), and positivity/liking (anova 3) across 3 groups (gay/lesbian, straight/heterosexual, and bisexual). For any that are significant, we will examine pairwise differences with Tukey posthoc tests[18] with Holm-Bonferroni corrections.

A repeated measures one-way anova revealed significant differences in participants' ratings of legitimacy across gay/lesbian, straight/heterosexual, and bisexual people, $F(1.74, 102.63) = 9.931$, $p < .001$, and planned post-hoc t-tests with Holm-Bonferroni corrections revealed that participants rated both gay/lesbian ($p < .001$) and straight/heterosexual people ($p = .02$) as significantly more real than bisexual people. However, there was no difference in perceived realness between gay/lesbian and straight/heterosexual people ($p > .05$).

A repeated measures one-way anova revealed significant differences in participants' ratings of stability across gay/lesbian, straight/heterosexual, and bisexual people, $F(1.56, 91.76) = 21.135$, $p < .001$, and planned post-hoc t-tests with Holm-Bonferroni corrections revealed that participants rated both gay/lesbian ($p < .001$) and straight/heterosexual people ($p < .001$) as significantly more stable than bisexual people. Additionally, straight/heterosexual people were perceived as more stable than gay/lesbian people ($p = .01$).

A repeated measures one-way anova revealed significant differences in participants' liking across gay/lesbian, straight/heterosexual, and bisexual people, $F(1.25, 73.91) = 7.791$, $p = .003$, and planned post-hoc t-tests with Holm-Bonferroni corrections revealed that participants liked straight/heterosexual people more than both bisexual people ($p = .009$) and gay/lesbian people ($p = 02$). However, post-hoc t-tests revealed no difference in participants' reported liking of gay/lesbian and bisexual individuals ($p > .05$).

[17]This is a post-exclusion N. Applying the same exclusion criteria as used in Study 1a resulted in the exclusion of one participant.

[18]Here and throughout, we run pairwise t-tests in lieu of our pre-registered Tukey tests (which require independent groups) because all data is paired within participants.

If we find support for hypotheses 1 and 3 (i.e. that bisexual people are perceived as less legitimate/socially real and liked less than both gay/lesbian and straight/heterosexual people at the pairwise level, we will run a secondary mediation analysis to test whether or not social realness is significantly mediated the extent to which bisexual people are evaluated more negatively than gay/lesbian and straight/heterosexual people. We will use a PROCESS macro in R and will enter group predictor, social realness/legitimacy as a mediator, and positivity/liking as an outcome.

Post-hoc t-tests revealed no difference in participants' reported liking of gay/lesbian and bisexual individuals. Thus, hypothesis 3 was not fully supported and a secondary mediation analysis was not run.

We will run 3 anovas looking at overall differences in legitimacy/social realness (anova 1), stability (anova 2), and positivity/liking across 3 groups (men, women, nonbinary people. For any that are significant, we will examine pairwise differences with Tukey pos-thoc tests[1] with Holm-Bonferroni corrections.
(These majority of these results are reported in the main manuscript.) However, non-focal planned post-hoc t-tests with Holm-Bonferroni corrections revealed no significant differences in ratings of men and women in the domains of legitimacy/social realness, stability, or liking (all $p \geq .05$).

Additionally, if we find support for hypotheses 4 and 6 (i.e. that nonbinary people are perceived as less legitimate/socially real and liked less than both men and women at the pairwise level, we will run a secondary mediation analysis to test whether or not social realness is significantly mediated the extent to which nonbinary people are evaluated more negatively than men and women. We will conduct a mediation to see whether or not social realness/legitimacy mediates the extent to which nonbinary people are evaluated more negatively than other groups. Again, we will use a PROCESS macro in R[19] and will enter group predictor, social realness/legitimacy as a mediator, and positivity/liking as an outcome.
We conducted a mediation analysis to examine whether the effect of gender condition on liking was mediated by perceived realness. The total effect of condition on liking was significant for both women in comparison to nonbinary people, $c = 31.5$, $p < .001$ (see Figure S1) and men in comparison to nonbinary people, $c = 25.8$, $p < .001$ (see Figure S2)

For both men and women, there were significant indirect effects of gender condition on liking through realness such that gender condition had a significant effect on realness ($a = 1.617$, $p < .001$ when comparing men to nonbinary people; $a = 1.586$, $p < .001$ when comparing women to nonbinary people), and realness had a significant positive effect on liking ($b = 8.85$, $p < .001$). The direct effect of gender condition on liking, controlling for realness, was $c' = 11.56$, $p < .05$ when comparing men vs. nonbinary people, and $c' = 17.46$, $p < .001$ when comparing women vs. nonbinary people, indicating partial mediation in both instances.

[19] Note that we did not use the PROCESS macro due to complications with multicategorical predictors, and instead implemented the mediation analysis in R by hand according to the steps outlined in Baron & Kenny, 1986.

Figure S1. Looking to see if perceived realness mediates negative evaluations of nonbinary people vs. women

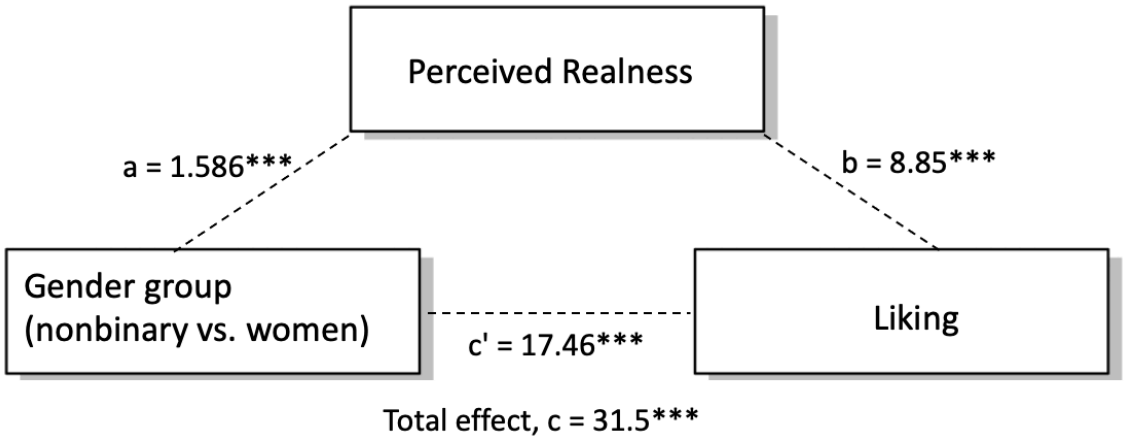
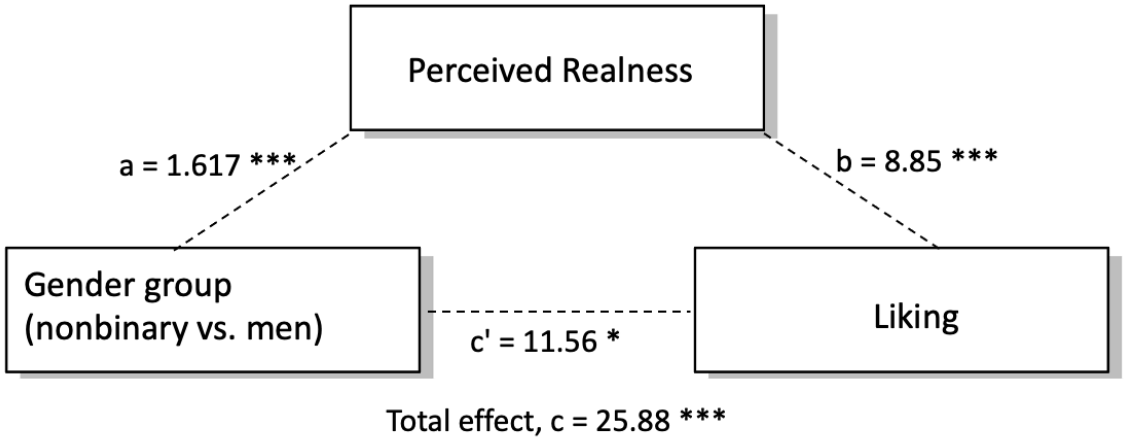



Figure S2. Looking to see if perceived realness mediates negative evaluations of nonbinary people vs. men




**Additional Information about Studies 1b & 1c[20]**

***Other preregistered analyses***

If we find that nonbinary people are perceived as both less socially real and are rated less favorably than the other groups in either comparison set, we will run a secondary mediation analysis to test whether or not social realness mediates the extent to which nonbinary people are evaluated more negatively than the other groups in the comparison set.

We found that nonbinary people were perceived as less socially real and liked less than both trans men and trans women (see analyses reported in the main manuscript). Therefore, we conducted a mediation analysis to examine whether the effect of gender condition on liking was mediated by perceived realness. The total effect of condition on liking was not significant for transgender men in

[20]All preregistered (main) analyses are reported in the main manuscript.

comparison to nonbinary people (see Figure S3), or transgender women in comparison to nonbinary people (see Figure S4; both *p*'s > .05).

For both transgender men and transgender women, there were significant indirect effects of gender condition on liking through realness such that gender condition had a significant effect on realness ($a = 0.757$, $p < .001$ when comparing transgender men to nonbinary people; $a = 0.852$, $p < .01$ when comparing transgender women to nonbinary people), and realness had a significant positive effect on liking ($b = 10.08$, $p < .001$). The direct effect of gender condition on liking, controlling for realness, was not significant ($p > .05$) when comparing transgender men vs. nonbinary people, or ($p > .05$) when comparing transgender women vs. nonbinary people, suggesting an indirect-only mediation (see Figure S3) and/or a suppression effect (see Figure S4).

Figure S3. Looking to see if perceived realness mediates negative evaluations of nonbinary people vs. transgender men

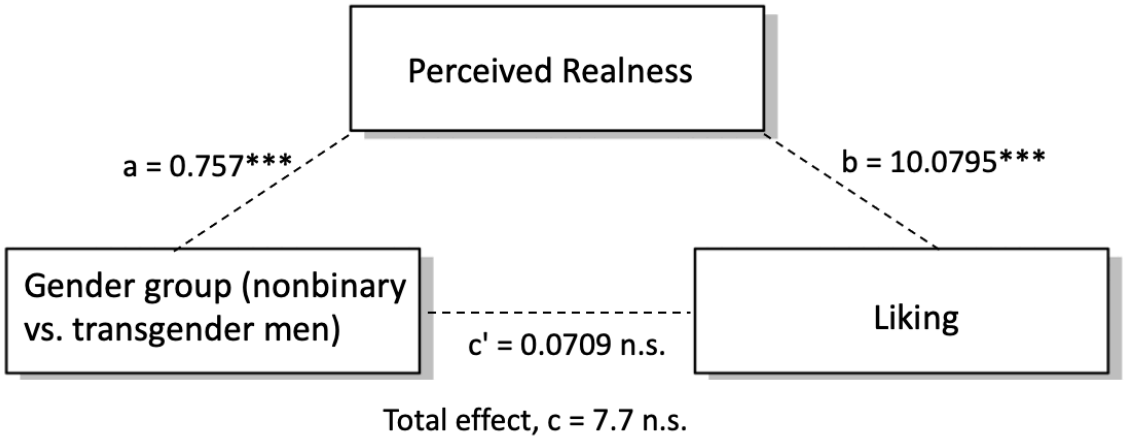
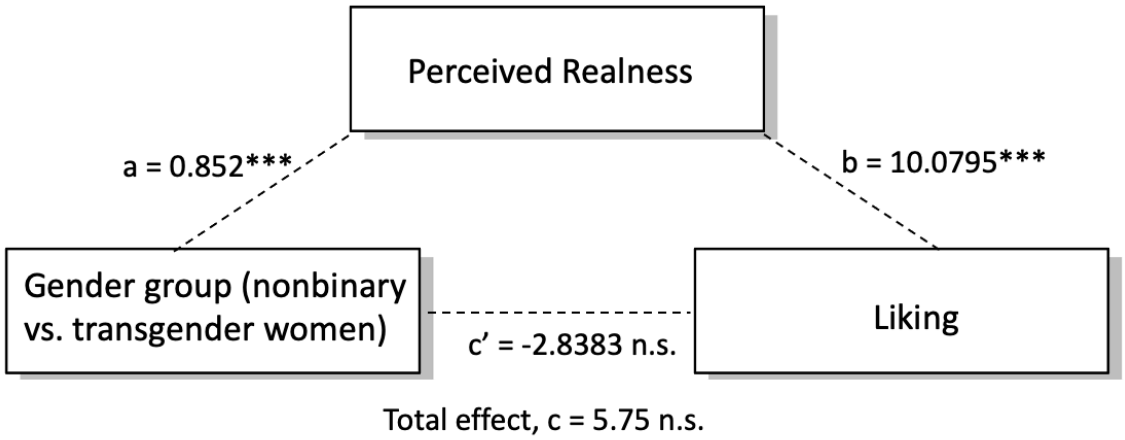



Figure S4. Looking to see if perceived realness mediates negative evaluations of nonbinary people vs. transgender women




If we find that nonbinary people are perceived as both less stable and are rated less favorably than the other groups in either comparison set, we will run a secondary mediation analysis to test whether or not

stability mediates the extent to which nonbinary people are evaluated more negatively than the other groups in the comparison set.

We found that nonbinary people were perceived as less stable and liked less than both trans men and trans women (see analyses reported in the main manuscript). Therefore, we conducted a mediation analysis to examine whether the effect of gender condition on liking was mediated by perceived stability. The total effect of condition on liking was not significant for transgender men in comparison to nonbinary people (see Figure S5), or transgender women in comparison to nonbinary people (see Figure S6; both *p*'s > .05).

For both transgender men and transgender women, there were significant indirect effects of gender condition on liking through stability such that gender condition had a significant effect on stability ($a = 0.93$, $p < .01$ when comparing transgender men to nonbinary people; $a = 1.012$, $p < .01$ when comparing transgender women to nonbinary people), and stability had a significant positive effect on liking ($b = 13.941$, $p < .001$). The direct effect of gender condition on liking, controlling for realness, was not significant ($p > .05$) when comparing transgender men vs. nonbinary people, suggesting an indirect-only mediation. When comparing transgender women vs. nonbinary people, the direct effect of gender condition on liking controlling for stability was significant ($c' = -8.358$, $p < .05$), suggesting stability partially mediates the relationship between gender condition and liking.

Figure S5. Looking to see if perceived stability mediates negative evaluations of nonbinary people vs. transgender men

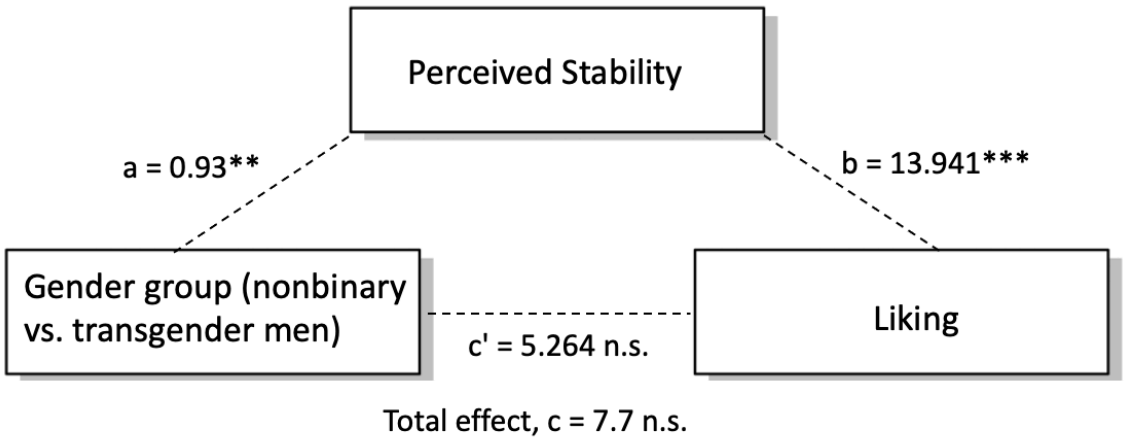
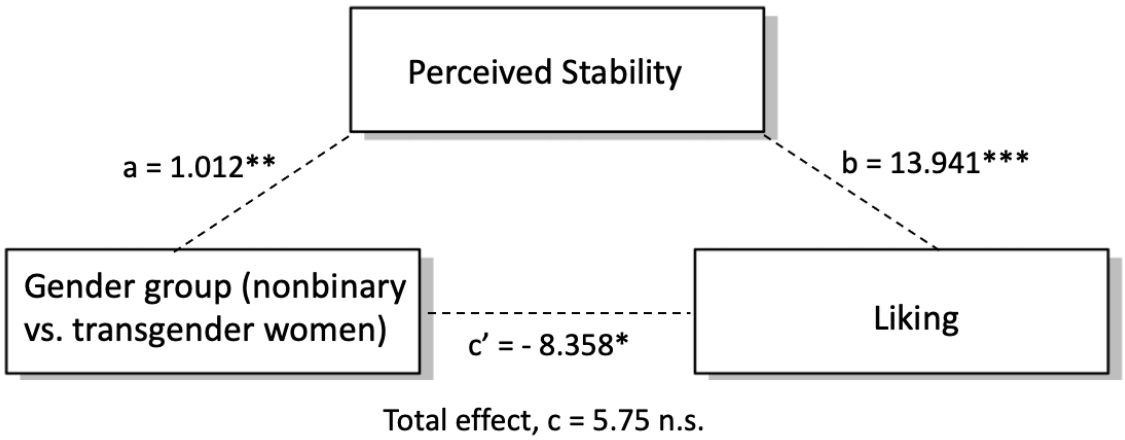


Figure S6. Looking to see if perceived stability mediates negative evaluations of nonbinary people vs. transgender women




**Additional Information about Study 2[1,21]**

***Other preregistered analyses***

In addition to our main analyses, we want to ask whether scores on the three main measures (liking, legitimacy, and stability) show differences for the three primary groups when they are treated as distinct groups (men, women, nonbinary). For each measure (liking, legitimacy, stability) we will run a one-way anova. If significant, we will examine pairwise differences with post hoc t-tests with Holm-Bonferroni corrections.

A repeated measures one-way anova revealed significant differences in participants' liking of men, women, and nonbinary people, $F(1.31, 231.83) = 110.78$, $p < .001$. Planned post-hoc t-tests with Holm-Bonferroni corrections revealed that participants liked both men and women significantly more than they liked nonbinary people, and liked women significantly more than men (all $p$'s $< .001$).

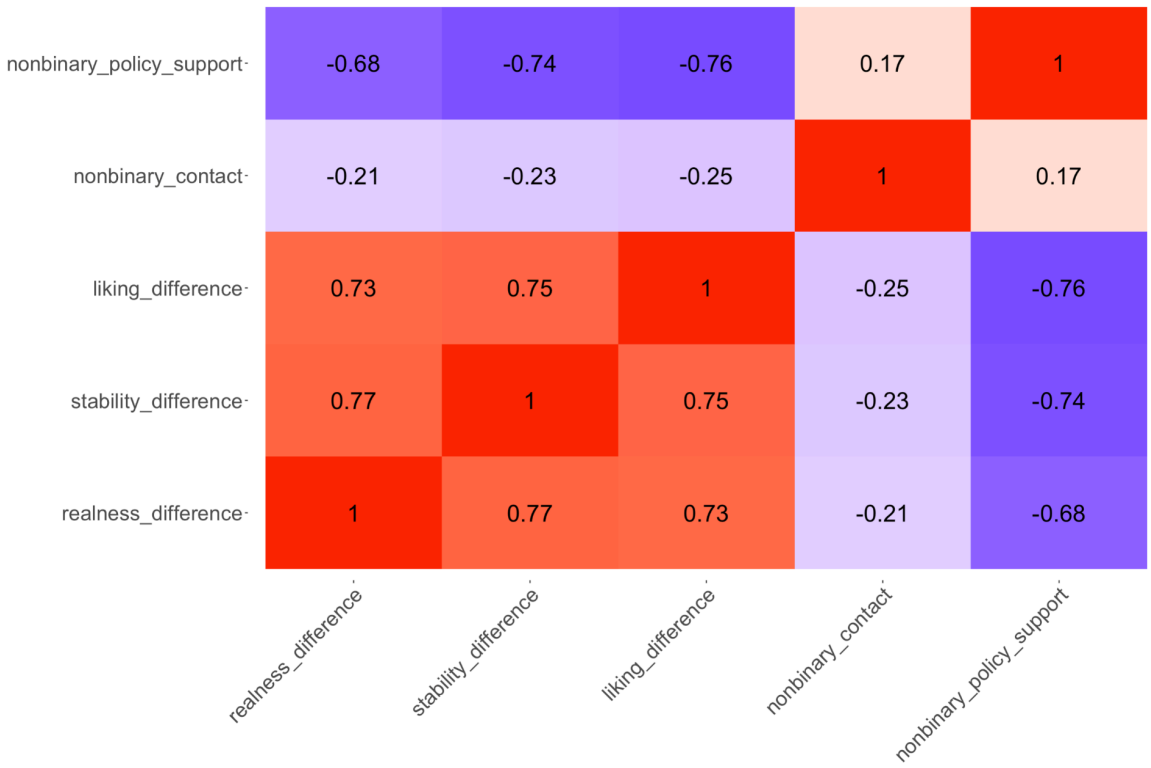
A repeated measures one-way anova revealed significant differences in participants' ratings of legitimacy across men, women, and nonbinary people, $F(1.14, 201.14) = 125.67$, $p < .001$, and planned post-hoc t-tests with Holm-Bonferroni corrections revealed that participants rated both men and women as significantly more real than nonbinary people (all $p$'s $< .001$). However, there was no difference in participants ratings of legitimacy between men and women ($p > .05$).

A repeated measures one-way anova revealed significant differences in participants' ratings of stability across men, women, and nonbinary people, $F(1.14, 201.27) = 244.65$, $p < .001$. Planned post-hoc t-tests with Holm-Bonferroni corrections revealed that participants rated both men and women as significantly more stable than nonbinary people (both $p$'s $< .001$). Additionally, participants rated women as significantly more stable than men ($p = .01$).

***Other analyses***

Figure S7. Correlations between nonbinary policy support, nonbinary contact, and biases against nonbinary people

[21] When describing the conditions of this study, our preregistration states that we "will test for order effects" However, we did not preregister any analyses that test for order effects, and thus none are included in this document.


nonbinary_policy_support
nonbinary_contact
liking_difference
stability_difference
realness_difference
-0.68
-0.74
-0.76
0.17
1
-0.21
-0.23
-0.25
1
0.17
0.73
0.75
1
-0.25
-0.76
0.77
1
0.75
-0.23
-0.74
1
0.77
0.73
-0.21
-0.68
realness_difference
stability_difference
liking_difference
nonbinary_contact
nonbinary_policy_support

**Cisgender people's perceptions of candidates for gender affirming surgery**

S. Atwood,[1,^] Kristina R. Olson,[1] & Thekla Morgenroth[2]

[1]Department of Psychology, Princeton University, Princeton, NJ 08540 USA.

[2]Department of Psychological Sciences, Purdue University, West Lafayette, IN 98105 USA.

S. Atwood: https://orcid.org/0000-0003-2889-9575

Kristina R. Olson: https://orcid.org/0000-0002-1111-544X

Thekla Morgenroth: https://orcid.org/0000-0001-9876-5017

Funding: This study was funded by Princeton University and a National Science Foundation Graduate Research Fellowship awarded to S. Atwood. Subsets of these data have been presented at the Annual Meeting of the Society for Personality and Social Psychology, the European Association of Social Psychology General Meeting, and at social psychology seminars at Princeton University and the University of Delaware.

Conflict of Interest: The authors declare that they have no conflicts of interest.

^Please direct correspondence concerning this article to: satwood@princeton.edu

**Abstract**

Access to gender affirming healthcare for trans and nonbinary people remains largely within the jurisdiction of cisgender people. Across five studies (N =1,336), we aim to investigate how cisgender people prioritize trans and nonbinary people for gender affirming healthcare. In studies 1a-1c, we find that cisgender people often (but not always) prioritize individuals who come out earlier (vs. later), have a gender expression that aligns with the surgery (ie, a feminine vs. masculine trans woman), and who have more binary (vs. nonbinary) identities. In studies 2 and 3, we investigate why cisgender people may prioritize these cases and find that cisgender participants rate binary trans individuals and people who come out at earlier ages as more likely to want surgery than nonbinary individuals and those who came out at later ages. We also find that cisgender people are more likely to recommend surgery candidates who say that they want surgery more. However, when two surgical candidates are matched for desire for surgery, we find that cisgender participants prioritize the binary transgender person over the nonbinary person for surgery access. Collectively these findings suggest a bias toward favoring binary candidates for surgery that may be partially explained by cisgender participants believing they want surgery more.

Keywords: gender affirming medical care, surgery recommendation, transgender nonbinary

In the past decade the number of trans and nonbinary people seeking gender-affirming healthcare has increased (Nolan et al., 2019; Agochukwu-Mmonu et al., 2022). However, research suggests that trans and nonbinary people face unique barriers within the healthcare system that make it difficult for them to access gender-affirming care (Houssayni & Nilsen, 2018; dickey et al., 2016; Bakko & Kattari, 2020; Kattari & Hasche, 2016; Lefevor et al., 2019; Kattari et al., 2019; Safer et al., 2016; Lerner & Robles, 2017; many more). One commonly reported barrier is the lack of availability of gender-affirming care providers, such that there are many more trans and nonbinary individuals who want gender-affirming care than there are qualified healthcare providers (Puckett et al., 2018; Cohen et al., 2020). We argue that when demand exceeds supply (as is the case with gender affirming healthcare in the United States), healthcare practitioners and other stakeholders (who are largely cisgender) must make decisions about who to prioritize for gender affirming care. In this work, we explore how cisgender people make decisions on who to prioritize for gender affirming care. Specifically, we ask whether cisgender people prioritize individuals who have binary (vs. nonbinary) identities, those who come out as transgender/nonbinary earlier in development, and those whose gender expression aligns with their surgery (e.g., those with a more feminine vs. masculine presentation for a vaginoplasty).

*Background & availability of gender affirming healthcare*

Gender affirming healthcare can be broadly defined as any interaction with the healthcare system that is responsive to the needs of trans, nonbinary, and/or gender diverse individuals (Loo et al.). For the purposes of this paper, we focus specifically on surgery (Berli et al., 2017; Akhavan et al., 2021). Prior research has found that access to gender affirming medical interventions is largely associated with improved mental health and well-being for trans and

gender diverse adults (Wernick et al., 2019; Nguyen et al, 2018; Swan et al., 2022). Indeed, research has shown that undergoing gender affirming medical interventions as part of transition may be life-saving for those who want it (Almazan & Keuroghlian, 2021; Hughes et al., 2021; Kidd et al., 2021).

Despite these largely positive consequences, gender-affirming care remains difficult to access. This difficulty arises from the simultaneous increase in demand and limited supply of providers. For instance, the American Society of Plastic Surgeons reported a 289% increase in gender-affirming surgeries for transgender men and a 41% increase among transgender women between 2016 and 2017 (Surgeons ASoP, 2017; Mousavian et al., 2024). A more recent study by Cohen et al. (2020) estimated that there are only 660 surgeons available to provide gender affirmation surgery in the USA, serving over 1 million transgender individuals. As a result, surgeons must be selective when deciding which surgeries to perform, and many people who want to pursue gender-affirming surgery are ultimately not able to do so (Sevelius & Jenness, 2017; Clark et al., 2018).

*Allocation and access to gender affirming healthcare*

Even in regions with greater availability of practitioners, the process of accessing gender-affirming requires support from individuals outside the surgery center. This is, in part, because the medical establishment treats transgender identity with a high degree of skepticism and requires an unusually high burden of identity 'proof' for trans and gender diverse people in order to access gender-affirming healthcare procedures like surgery. This may include (but is not limited to) formal diagnosis of Gender Dysphoria and/or letters from mental health professionals signing off on the legitimacy their gender identity in order to be eligible to pursue other forms of medical gender-affirming healthcare (Coleman et al., 2022; Cavanaugh et al., 2016; Ashley,

2019). Another common barrier to accessing gender affirming surgery is cost. Gender affirming surgery typically costs tens of thousands of dollars (Downing et al., 2022; Chu et al., 2024), and thus is not feasible for most surgery seekers to pay for out of pocket. As a result, trans and nonbinary surgery seekers must rely on individuals working at health insurance companies to approve their medical claims to be able to afford the cost of gender affirming surgery. Finally, many gender-affirming procedures are considered major surgery and require social support. This means that surgery seekers rely on family and friends to drive them to and from the hospital, and help with basic tasks (e.g. getting dressed, eating, bathing, aftercare etc.) during recovery (Puckett et al., 2018). Thus, at every step of the process, transgender and nonbinary people are often forced to rely on cisgender people in their lives to 'recommend' and/or give them social/financial support required in order to successfully obtain gender affirming surgery.

*Cisgender people's considerations when allocating gender affirming healthcare*

Given the impact of cisgender people's decisions in allocating healthcare to trans and nonbinary individuals, it is crucial to understand the factors cisgender individuals consider when making these decisions. One possibility is that cisgender people make decisions about gender affirming surgery based on their beliefs about who would commonly seek or want this surgery. We test three dimensions that we predicted cisgender people might use to make these judgments. First, we test whether participants believe individuals with a binary identity (operationalized in these particular studies as trans women) would be a better candidates for or want surgery more than individuals with a nonbinary identity (and who were assigned male at birth). This is because transgender people have historically been the most common group to seek gender-affirming surgery (Nolan et al., 2019) and so may be especially salient in participants' minds when thinking about candidates for gender affirming surgery. Second, we test whether participants

believe that someone who reports having felt that they were transgender or nonbinary since early in life will be a better candidate or want surgery more than someone who comes out as transgender or nonbinary later in life. This is reflected in stereotypes in popular culture (e.g. narratives about being "born in the wrong body"; Bettcher, 2014) and also aligns with literature on trans identity development which converges around the idea that individuals who assert their identities early, consistently, and often are especially good candidates for gender-affirming healthcare (Forcier & Haddad, 2013; Steensma et al., 2013; Levitt & Ippolito, 2014; Johnson, 2015). Finally, we test whether or not participants think that surgery seekers whose gender presentation is different from the typical gender presentation associated with their assigned sex at birth (e.g. a trans woman with stereotypically feminine gender expression) would be a better candidate than someone whose gender expression coincides with their assigned sex at birth (e.g. a trans woman with an especially masculine gender expression). Cisgender participants might come to prioritize these candidates because these aspects of a person's gender identity, history, and/or presentation align with their ideas about who wants surgery, possibly driven by popular representations of trans identity which often focus on trans individuals who embody cisgender/heteronormative ideals and reinforce dominant, binary gender ideologies (Glover, 2016; Van den Berg & Marinus, 2017).

*Current Research*

In a series of five online studies, we examine how cisgender people prioritize binary transgender and nonbinary people for gender affirming surgery. We test three hypotheses: that they will prioritize transgender women  over nonbinary candidates assigned male at birth (Studies 1a-1c, and 3), that they will prioritize individuals with earlier ages of first gender identification (Studies 1a-1c), and will prioritize individuals whose gender expression aligns less

stereotypically with their assigned sex at birth (Studies 1a and 1b). We then ask whether these prioritizations might relate to cisgender people's perceptions of who most wants surgery (Studies 2 & 3).

To ensure transparency and openness, all preregistrations, materials, data, and code for all studies are available on the OSF project page (see https://osf.io/z8x9y/?view_only=83da42afcbaf4b4b85a8a996d380405b). Studies 1a-1c and Study 3 were preregistered (https://osf.io/gj4xp/?view_only=25c822261d5f4503bc1649d4ba057159) and include all exclusion criteria used in the current manuscript. In the main text, we report a subset of the pre-registered analyses as well as additional analyses that go beyond the scope of the preregistration. However, all preregistered analyses are fully reported in the manuscript or the Supplementary Information (see https://osf.io/wqds8/?view_only=9474395efd65454586e0097ffac6640b).

Finally, all studies involving human participants were done in accordance with the ethical standards of the Human Subjects Division Institutional Review Board at [masked for peer review] and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Additionally, all participants provided informed consent prior to participating in the study.

## Studies 1a - 1c

In the first set of studies, we test how cisgender people prioritize individuals for gender affirming surgery using nearly identical methodology with the exception of how details varied within individual stimuli. All three of these studies were pre-registered (see https://osf.io/76cyz?view_only=8ad7c065297c4a69af7a4dc1b02024b5 for Study 1a;

https://osf.io/q7569?view_only=8ad7c065297c4a69af7a4dc1b02024b5 for Study 1b; and see https://osf.io/7sujn?view_only=8ad7c065297c4a69af7a4dc1b02024b5 for Study 1c). We combine the presentation of these studies here for ease of comprehension of the overall pattern of results and to reduce redundancy.

*Method*

*Participants*. In Study 1a, a convenience sample of 209 participants completed the study on CloudResearch in Spring 2022. We preregistered our sample size and exclusions in advance of data collection, aiming for 199 eligible participants to give us more than 80% power to detect an effect of d = .2 or greater in our planned post-hoc t-test analyses. We used the same pre-registered exclusion criteria for all pre-registered studies (Studies 1a-1c and Study 3). Specifically, we excluded data from participants who failed attention checks (those who admit to not paying attention, and those who fail the comprehension check question), and those who took an unreasonably long or impossibly short amount of time to complete the survey (>4x the mean survey completion time and < .25 the mean survey completion time, respectively). Also, because we were interested in cisgender people's beliefs, we excluded any participants who do not self-identify as cisgender. Together, these criteria excluded 5 individuals, which left us with 204 eligible participants in Study 1a (see Table 1 for demographic information).

In Study 1b, a convenience sample of 223 participants completed the study on CloudResearch in Spring 2022. Like before, we preregistered our sample size in advance of data collection, this time aiming for 202 eligible participants to give us more than 80% power to detect a significant interaction effect of gender by prototypicality (as based on covariance estimates from Study 1a). Using the same exclusion criterion as in Study 1a, we excluded 21

individuals, which left us with 202 eligible participants in study 1b (see Table 1 for demographic information).

In Study 1c, a convenience sample of 793 participants completed the study on CloudResearch in Summer 2022. We preregistered our sample size and exclusion criteria in advance of data collection, aiming for 640 participants to give us 80% power in our preregistered mediation analysis[22] (See SI). Based on our pre-registered exclusion criteria (see Study 1a above) we excluded 163 individuals[23], which left us with 630 eligible participants in Study 1c (see Table 1 for demographic information).

Table 1. Demographic information for all studies

| Study Number | Study 1a | Study 1b | Study 1c | Study 2 | Study 3 |
|---|---|---|---|---|---|
| Current Gender | | | | | |
| Woman | 53% | 52% | 63% | 49% | 50% |
| Man | 47% | 48% | 37% | 51% | 50% |
| Race/Ethnicity | | | | | |
| Asian | 5% | 3% | 3% | 9% | 6.5% |

[22] We conducted a power analysis based on guidelines set out by Schoemann and colleagues (2017). First we estimated a range of possible correlations between the variables in our parallel mediation model based on data from our previous studies, and estimates from related literature (Weber et al., 2002; Tee & Hegarty, 2006; Roberts et al., 2017; Wilton et al., 2018; Morgenroth et al., 2021). Then, we used a Monte Carlo power analysis simulation to estimate power at a range of possible N's. Based on our simulations, we aim to collect 320 participants for each of our models (so 640 participants total) in order to give us at least 80% power using our mediation models.

[23] Unlike previous studies where we recruited participants directly via crowd-sourcing platform, participants in Study 3 were recruited via online panel managed by a 3rd party, which are known to have poorer data quality and higher error rate on attention check questions (Zhang & Gearhart, 2020).

| | | | | | |
|---|---|---|---|---|---|
| Black | 6% | 8% | 18% | 6% | 11.5% |
| Latinx | 1.5% | 2% | 5% | 8% | 3% |
| Native American/ Alaska Native | . | . | 1% | 1% | < 1% |
| Multiracial | 6% | 8% | 6% | 12% | 6% |
| White | 81% | 78% | 66% | 64% | 72% |
| Jewish | . | 1% | < 1% | . | . |
| Other/Did Not Disclose | .5% | . | < 1% | . | < 1% |
| Age | | | | | |
| Mean (in years) | 43.80 | 41.76 | 41.80 | 40.99 | 41.81 |
| SD (in years) | 12.00 | 12.44 | 15.30 | 13.84 | 14.03 |
| Total N (per study) | 204 | 202 | 630 | 101 | 199 |

*Procedure.* In these studies, participants read about individuals who were all interested in undergoing surgery as part of their gender affirming medical transition. After reading about each individual, participants were asked to share their opinion about whether or not the individual should be approved to have surgery (our main variable of interest) using a sliding scale of 0 (definitely would not recommend) to 100 (would definitely recommend). Additionally, participants were told that the surgeon was busy and did not have time to perform surgery on

everyone who was interested, and so their recommendations would be taken into account in order to figure out which surgeries to perform and which not to. In study 1a and 1b, participants saw 14 vignettes and in study 1c participants saw only one vignette. After providing their recommendations, participants completed a brief attention check and demographic questionnaire. All data were collected via Qualtrics.

*Vignettes*. All individuals in the vignettes were assigned male at birth and were between the ages of 18-31. Every vignette included basic details about the individuals' life (e.g. gender identity, home city, hobbies, and occupation; see Table 2). Individuals in the vignettes varied along three central dimensions: gender identity (applies to Studies 1a – 1c; where individual was depicted as either a binary transgender woman or a nonbinary person assigned male at birth), target gender expression (applies to Studies 1a and 1b only; where individual was depicted as having traditionally feminine gender expression or traditionally masculine gender expression), and age of identification (applies to Studies 1a-1c; where individual was depicted as having identified as their gender early in childhood or within the past year (Studies 1a & 1b) or depicted as having identified as their gender early in childhood or in adolescence (Studies 1c only). For gender identity, we expected vignettes about transgender women to be prioritized relative to those about nonbinary people assigned male at birth. For gender expression, we expected vignettes about traditionally feminine people to be prioritized relative to those about traditionally masculine people. For age of identification, we expected people who identified as their gender early in childhood to be prioritized relative to people who identified their gender within the past year (studies 1a & 1b), but had no directional prediction about how people who identified as their gender early in childhood would be prioritized relative to people who identified their gender in adolescence (study 1c).

In studies 1a and 1b, each participant saw four vignettes about a transgender woman (feminine expression, masculine expression, early identification, and late identification), and four parallel vignettes about a non-binary person assigned male at birth[24]. Each vignette was therefore defined by individuals' gender identity and one non-gender-identity dimension of expected prioritization (gender expression or age of identification). In study 1a, individuals' demographic profiles (e.g., name, age, occupation, etc.) were associated with one gender identity (i.e., Jules the 25-year old lawyer only appeared as a trans woman, never as a non-binary person), while in study 1b, each profile was evenly associated with each gender identity group.

In study 1c, we simplified our design and did not manipulate gender expression. We also changed our age of first identification variable to depict individuals who came out in either childhood or in adolescence (Table 2). Study 1c therefore had four trial types (gender identity x age of identification), with individuals'' demographic profiles balanced across the trials. Each participant was randomly assigned to evaluate only one vignette.

Table 2

| Category | Predicted to be prioritized | Predicted to be deprioritized |
|---|---|---|
| Gender identity (all studies) | Transgender | Nonbinary |
| Gender expression (studies 1a & 1b only) | Feminine Expression | Masculine Expression |
| Age of first gender identification (studies 1a and 1b only) | In Childhood | Recently |
| Age of first gender identification (study 1c only) | Compared individuals who came out in adolescence vs. in childhood | |

[24] Participants also saw six "filler" vignettes; that we do not analyze this data for this manuscript

Table 2. List of characteristics embedded within vignettes by study.

Table 3

| | Trans Woman | Nonbinary |
|---|---|---|
| First Identification | Jules is a 25 year old **trans woman** from the suburbs outside of Philadelphia, PA. Growing up, **she** was actively involved with Boy Scouts of America and earned the rank of Eagle Scout at age 17. She enjoys spending weekends thrift shopping with her friends. She graduated a few years ago with a degree in international relations from Villanova University. She is working at a small law firm that focuses on international corporate law. **Around age 3, she remembers feeling certain that she was a girl** and she first told her parents in third grade. Starting in middle school, she came out to her friends and family. She feels ready to have a vaginoplasty now. She has average health and a sibling who would be available for post-op support. | Jules is a 25 year old **nonbinary person** from the suburbs outside of Philadelphia, PA. Growing up, **they** were actively involved with Boy Scouts of America and earned the rank of Eagle Scout at age 17. They enjoy spending weekends thrift shopping with their friends. They graduated a few years ago with a degree in international relations from Villanova University. They are working at a small law firm that focuses on international corporate law. **They did not think much their gender identity until about a year ago, when they remember feeling for the first time like they were nonbinary.** Since then, they have come out to the majority of their family and friends as nonbinary and they want to have a vaginoplasty now. They have average health and a sibling who would be available for post-op support. |
| Gender Expression | Hadley is a 23 year old **trans woman** from Savannah, GA. Raised as a boy, **she** followed her father's footsteps and served several years in the Peace Corps after earning her GED. More recently, she started pursuing an undergraduate degree in plant systematics from Emory University and hopes to one day travel the world classifying new species of plants. On a typical day, you might find her with **her long hair up in a bun, wearing a full face of makeup, along with** | Hadley is a 23 year old **nonbinary person** from Savannah, GA. Raised as a boy, **they** followed their father's footsteps and served several years in the Peace Corps after earning their GED. More recently, they started pursuing an undergraduate degree in plant systematics from Emory University and hopes to one day travel the world classifying new species of plants. On a typical day, you might find them with **their short hair gelled back dressed in their favorite pair of cargo shorts and a polo shirt.** In their free time |

| | | |
|---|---|---|
| | **her favorite skirt, blouse, and jewelry.** In her free time she enjoys **trying out new recipes or working on one of her many crocheting projects**. Additionally, she has come out to the majority of her family and friends as trans and feels ready to have a vaginoplasty now. She is healthy and have the support of her family who would be available to care for her post-op. | they enjoy **going on fishing trips and spending time out in their garage doing one of their many woodworking projects**. Additionally, they have come out to the majority of their family and friends as nonbinary and feel ready to have a vaginoplasty now. They are healthy and have the support of their family who would be available to care for them post-op. |

Table 3. Sample vignettes used in studies 1a-1c.

*Analytic Approach*

We ran a series of regression models to predict likelihood of surgery recommendation from gender identity, gender expression, and age of gender identification. Across all models and studies, we contrast-coded three fixed effect predictors (gender identity, gender expression, and gender identification), with characteristics that were expected to be prioritized contrast-coded as 0.5, and those expected to be deprioritized contrast-coded as -0.5. Thus, transgender woman (expected to be prioritized in the domain of gender identity), traditionally feminine expression (expected to be prioritized in the domain of gender expression), and early identification (expected to be prioritized in the domain of gender identification) were all contrast-coded as 0.5, while nonbinary person (expected to be deprioritized in the domain of gender identity), traditionally masculine expression (expected to be deprioritized in the domain of gender expression), and recent identification (expected to be deprioritized in the domain of gender identification) were all contrast-coded as -0.5. In study 1c we had no predictions about prioritization based on gender identification, so we contrasted early identification [contrast-coded as 0.5] (to be consistent with Studies 1a and 1b) and adolescent identification

[contrast-coded as -0.5]. For regression models with repeated measures (i.e. studies 1a and 1b), we additionally included a random intercept for participant[25].

Across all studies, we report a combination of preregistered and exploratory models, focusing on a subset of the coefficients from both. In our preregistered models, we include all data from each study and examine whether support for surgery recommendation is predicted by gender identity, other surgery-relevant dimensions (i.e. the combined effect of gender expression and gender identification) and/or their interaction between gender identity and other surgery-relevant dimensions. In the manuscript, we report the coefficients related to the main effect of gender identity as well as the interaction between gender identity and other surgery-relevant dimensions (gender expression and/or age of first gender identification). In our exploratory models, we considered gender expression and age of identification as separate factors (rather than combined, as in our preregistered analyses). We used the same model structure as in the preregistered analyses, but subset the trials so that we were looking only at trials defined with respect to gender expression (feminine vs. masculine) or age of first identification (studies 1a & 1b: childhood vs. recently; study 1c: childhood vs. adolescence). In the manuscript, we report coefficients related to the main effect of gender expression and age of first gender identification, respectively. Full outputs of all models presented in the manuscript along with some additional preregistered models not presented in the main paper are included in full the Supplementary Information. Additionally, we include model-based follow-up tests after each regression such that our follow-up test methodology is consistent between the different studies (i.e., those with and without repeated measures)[26].

[25] Our preregistration incorrectly stated that study 1c analyses would include a random intercept for participant; this is not possible because each participant evaluated only one vignette.
[26] We had initially preregistered follow-up t-tests, but instead present model-based contrasts. All initially t-tests are reported as initially preregistered in the Supplementary Information.

All analyses were conducted using R [Version 4.3.2] software (R Core Team, 2023) via RStudio [Version 2023.09.1+494]. Analyses were performed using the following packages: tidyverse (v2.0.0, Wickham, 2023), knitr (v1.45, Xie, 2023), psych (v. 2.3.9, Revelle, 2023), lme4 (v1.1-35.1, Bates et al., 2024), lmerTest (v. 3.1-3, Kuznetsova et al., 2020), table1 (v. 1.4.3, Rich, 2023), sjPlot (v. 2.8.15, Lüdecke, 2023), and emmeans (v. 1.10.1, Lenth, 2024).

*Results*

Our preregistered models found a main effect of gender identity – participants were more likely to recommend trans individuals than nonbinary individuals for surgery (Study 1a; $b$ = 13.36, $SE$ = 1.05, $p < .001$; Study 1b; $b$ = 12.99, $SE$ = 1.09, $p < .001$; Study 1c; $b$ = 12.17, $SE$ = 2.55, $p < .001$; see Figures 1 and 2). Additionally, exploratory models revealed main effects such that participants were more likely to recommend stereotypically feminine than stereotypically masculine individuals (Study 1a; $b$ = 6.69, $SE$ = 1.31, $p < .001$ and Study 1b; $b$ = 8.15, $SE$ = 1.47, $p < .001$) and those who identified in childhood rather than recently (Study 1a; $b$ = 19.89, $SE$ = 1.41, $p < .001$ and Study 1b; $b$ = 22.19, $SE$ = 1.52, $p < .001$), but not those who identified in childhood rather than adolescence ($p > .05$; see Figure 2).

Returning to the preregistered models, we found that gender expression and age of identification interact with gender identity in Studies 1a and 1b (Study 1a; $b$ = 8.50, $SE$ = 2.09, $p < .001$ and Study 1b; $b$ = 4.63, $SE$ = 2.18, $p$ = .034, also see Figure 2) but not in Study 1c ($p > .05$). Post-hoc model-based analyses from Studies 1a and 1b revealed that transgender people and nonbinary people who had an earlier identification and/or more feminine gender presentation were more likely to get recommended for surgery (transgender women – Study 1a: $t(1425)$ = 11.84, $p < .001$; Study 1b: $t(1411)$ = 11.35, $p < .001$; nonbinary people – Study 1a; $t(1425)$ =

6.105, $p < .001$; nonbinary people – Study 1b; $t(1411) = 8.35$, $p < .001$; also see Figure 1).

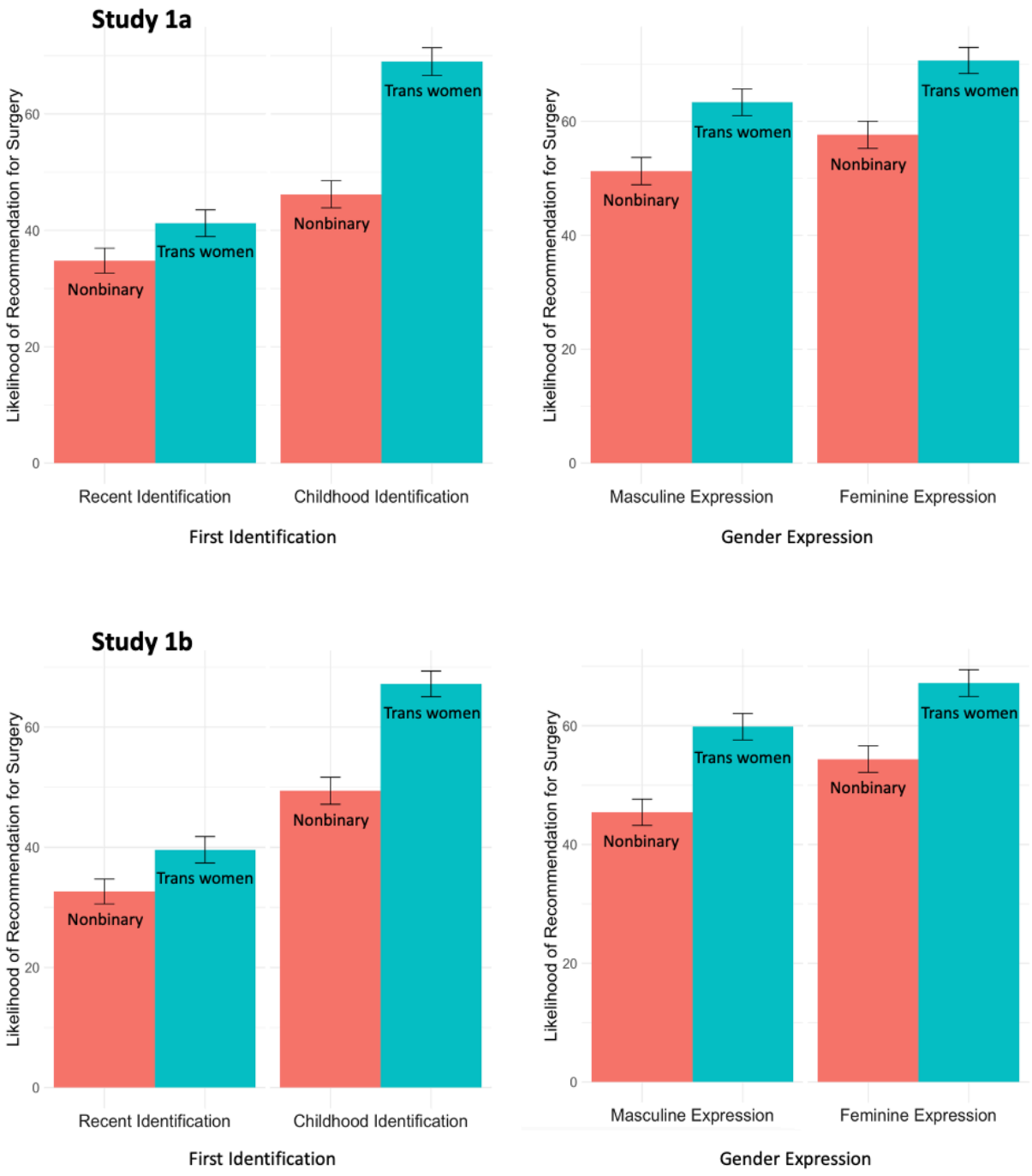



Figure 1. Likelihood of recommendation for surgery in Studies 1a (top) & 1b (bottom). In both studies, recommendation for gender-affirming surgery was higher for trans > nonbinary (all panels), childhood > recent identification (left-hand panels), and feminine > masculine presentation (right-hand panels)

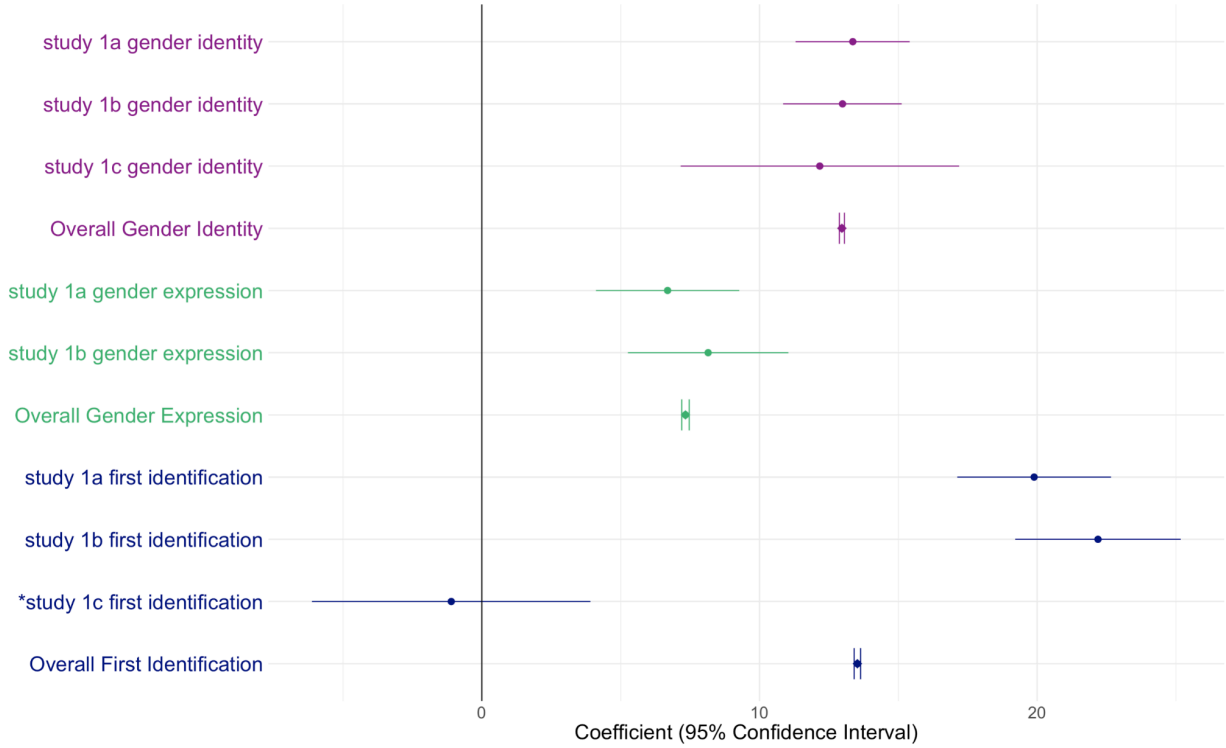



Figure 2. Forest plot summarizing strength of gender identity, gender expression, and first identification in studies 1a-1c. All study-specific coefficients reflect the main effect of each category within a study; errorbars are 95% CI. Overall coefficients represent weighted averages of each category across studies. *Study 1c compared likelihood of recommendation for surgery candidates who first identified their gender in childhood vs. adolescence (as opposed Studies 1a and 1b, which compared those who first identified their gender in childhood vs. within the last year).

*Discussion*

When it comes to accessing gender affirming surgery, we find that across these studies, cis people are overall more likely to recommend binary trans people than nonbinary people for gender affirming surgery. Additionally, cisgender people were more likely to recommend those who identified as trans/nonbinary earlier (at least by adolescence) and those who had a more feminine presentation.

## Study 2

We were next interested in understanding why participants might have prioritized binary (vs. nonbinary people), those who identified as trans/nonbinary earlier (vs. later), and those who

had more feminine (vs. masculine) presentations. Specifically, we explore the possibility that cisgender people think these groups have a higher desire for surgery. For exploratory purposes we expanded beyond (just) the dimensions we had previously examined to gain a broader understanding of what features of a candidate cisgender people may prioritize in making these decisions.

*Method*

*Participants*. In this study, a convenience sample of 113 participants completed the study on Prolific Academic in Summer 2023. Since this was a fully exploratory study, we did not preregister our sample size or exclusion criteria. We aimed for 100 participants and excluded anyone who admitted to to not paying attention, resulting in the exclusion of 12 individuals and leaving us with a final sample of 101 participants (see Table 1 for demographic information).

*Procedure*. In order to get a better sense of what participants thought was important when considering somebody for gender affirming surgery, we presented participants with a list of characteristics possibly relevant to a person seeking gender affirming surgery and asked participants to rank the extent to which each characteristic should be considered on a sliding scale from 0 (should definitely not be considered) to 100 (should definitely be considered) when deciding whether or not to recommend somebody to gender affirming surgery. The characteristics mirrored relevant information in the vignettes in our previous studies, but also went beyond those characteristics. We asked about the relevance of a person's current gender identity, a person's age when they first realized they were gender diverse, the amount of time a person has been living as their current gender identity, the extent to which a person's gender expression aligns with their current gender, how certain a person is that they want surgery, and others (for full list see Supplementary Information). Additionally, to get a sense of the types of

people participants thought wanted gender affirming surgery the most, we asked participants to rate the likelihood that people with various characteristics (e.g. a person who is transgender, a person who is nonbinary, a person who has just recently decided that they want surgery, a person who has been waiting to get surgery for a long time, a person who first realized their gender in early childhood, a person who first realized their gender in adolescence, etc.) would want gender affirming healthcare on a scale from 0 (very unlikely to want gender affirming surgery) to 100 (very likely to want gender affirming surgery; for full list of characteristics, see Supplementary Information).

*Results*

We ran a series of exploratory one-sample t-tests comparing participants' ratings to a neutral value of 50 to see which characteristics (if any) participants rated as particularly important or unimportant when determining whether or not to recommend a person for gender affirming surgery. In line with our previous findings, we found that participants thought it was more important than neutral to consider the amount of time a person has been living as their current gender identity ($t(100) = 4.21$, $p < .001$), and a person's age when they first realized they were gender diverse ($t(99) = 2.10$, $p = .04$). Conversely, participants thought it was less important than neutral to consider the extent to which a person's gender expression aligns with their current gender ($t(99) = -2.62$, $p = .01$), and reported that a person's current gender identity was neither important nor unimportant to consider when determining whether or not to recommend a person for gender affirming surgery ($p = .95$; see Figure 3 and see Supplementary Information).

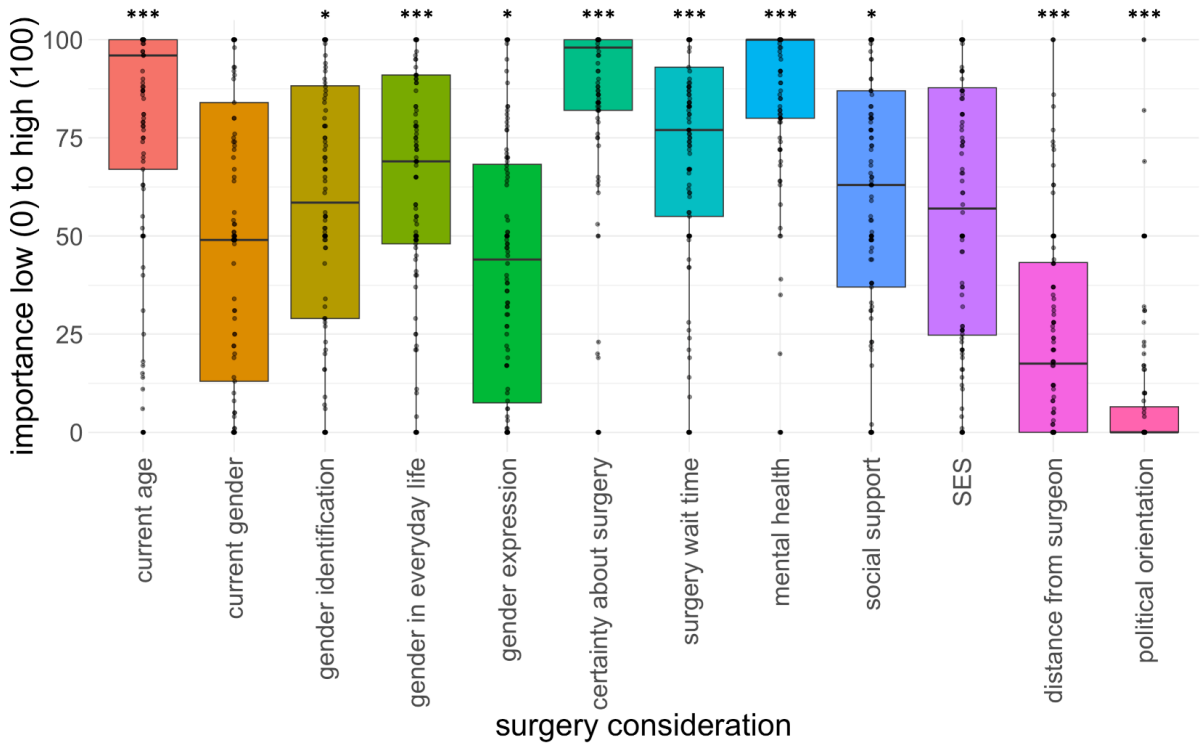



Figure 3. Importance of potential surgery considerations in comparison to a neutral value of 50 (*** denotes $p < .001$; ** denotes $p < .01$; * denotes $p < .05$)

Additionally, we ran a second series of exploratory one-sample t-tests (again comparing participants' ratings to a neutral value of 50) to see which characteristics participants believed were associated with wanting or not wanting gender affirming surgery. We found that even though participants did not think that a surgery seeker's gender identity was important to consider when recommending gender-affirming surgery (as noted above), they reported beliefs that transgender people were overwhelmingly more likely to want gender-affirming surgery ($t(100) = 15.21$, $p < .001$) while nonbinary people were less likely to want gender-affirming surgery ($t(100) = -2.33$, $p = .02$; also see Figure 4). Additionally, we found that participants rated individuals who just recently decided that they want surgery as well as those who had wanted surgery for a long time as both significantly more likely (than neutral) to want surgery ($t(100) = 4.00$, $p < .001$; and $t(100) = 22.76$, $p < .001$, respectively). A follow-up paired samples t-test

directly comparing ratings between those two groups revealed that participants rated those who had wanted surgery for a long time as significantly more likely to want surgery than those who had recently decided they wanted surgery ($t(100) = 11.39$, $p < .001$). Finally, we found that participants rated individuals who first realized their gender identity in early childhood, adolescence, and early adulthood as all significantly more likely (than neutral) to want surgery ($t(100) = 11.47$, $p < .001$; $t(100) = 11.04$, $p < .001$; and $t(100) = 10.76$, $p < .001$, respectively). However, a follow-up one-way within-participant anova directly comparing ratings between the three groups revealed no difference in likelihood of wanting surgery between groups ($F(2,200) = 2.22$, $p > .05$; see Figure 4 and see Supplementary Information).

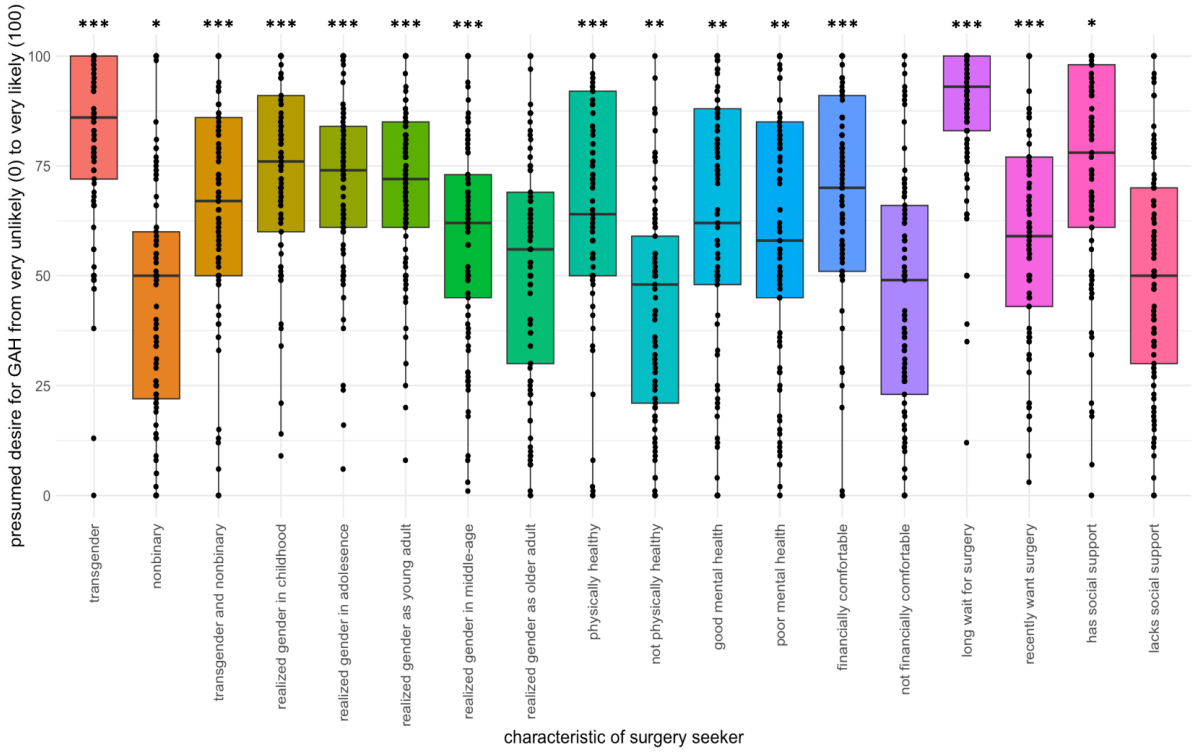



Figure 4. Presumed desire for gender affirming healthcare for surgery seekers with various characteristics in comparison to a neutral value of 50 (*** denotes $p < .001$; ** denotes $p < .01$; * denotes $p < .05$)

*Discussion*

Combined with results from Study 1, our first set of results from Study 2 present a contradiction. In Study 1, participants' recommendations varied by gender identity, consistently prioritizing trans people over than nonbinary people for surgery. In contrast, participants in Study 2 did not rate gender identity as particularly important when considering who should be prioritized for gender affirming surgery. However, the second set of results from Study 2, where we find that participants believed that trans people are much more likely to want gender-affirming surgery than nonbinary people, provides an explanation for these contrasting findings. Perhaps it is not participants' own ideas of what is important to consider, but rather participants' beliefs about how strongly different people want surgery that is driving decisions about who should be prioritized for surgery. In sum, the combined results of Studies 1-2 provide preliminary support for the possibility that cisgender people may rely on their beliefs about who wants surgery more to inform their own judgments of who to prioritize for surgery. In our final study, we will explore this possibility with regard to the binary transgender vs. nonbinary distinction.

## Study 3

In a final study we sought to better understand the binary vs. nonbinary distinction that participants made in all previous studies. While the results of Study 2 suggested that people think nonbinary people are less likely to want surgery, Study 3 investigated whether this assumption could be overridden by explicit information matching surgical candidates who were transgender vs. nonbinary on their stated interest in surgery. We described binary transgender and nonbinary candidates who either 'really wanted' surgery vs. 'wanted' surgery. In addition to rating candidates individually, we also asked participants who they would choose if asked to select just

one out of four candidates for surgery. If the tendency to favor binary transgender candidates is driven (exclusively) by a bias to see them as wanting surgery more, then providing evidence of equal interest should eliminate that prioritization.

*Method*

*Participants*. In this study, a convenience sample of 204 participants (See https://osf.io/qb6te?view_only=8ad7c065297c4a69af7a4dc1b02024b5) completed the study on Prolific Academic in Spring 2024. We preregistered our sample size and exclusion criteria in advance of data collection, aiming for 199 participants to give us 80% power to detect an effect size of $d = .2$ or greater in our planned post-hoc paired t-tests. Based on our pre-registered exclusion criteria (see criteria used in Study 1a) we excluded 5 individuals, which left us with 199 eligible participants in Study 2 (see Table 1 for demographic information).

*Procedure*. In this within-subjects study, participants viewed Surgery Readiness Forms, which were ostensibly completed by individuals interested in undergoing gender-affirming surgery. These forms contained basic demographic information (including name, age, pronouns, gender identity, employment status, etc.) as well as information specifically relevant to getting surgery (e.g. if they had a people to take them to and from the surgery, allergies, funding source, how much they wanted the surgery, etc.; see Supplementary Information) for example surgery readiness form). For the purposes of this study, the two key pieces of information on the form were the person's gender identity (listed as either a Transgender Woman or Nonbinary) and their desire for surgery/perceived importance of surgery (both listed as either 6-7 or 9-10 on a scale of "Not at all"/0 to "Very much"/10). Candidates with 'high' desire were those with an average across desire and perceived importance of 9.5 (one marked as 9 and one marked 10) and candidates with a 'regular' desire had an average of 6.5 across desire and perceived importance

of surgery (with one marked as 6 and one marked 7). Demographic profiles were randomized across gender identity and surgery desire, and participants saw one binary trans and one nonbinary person with each level of desire. The forms were presented one at a time in a random order. After seeing each form, participants were asked to share their opinion about whether or not the individual should be approved to have surgery (the same DV as in previous studies). Finally, after viewing all four forms individually, the participants were presented with all four forms at once and were asked to choose which individual they would be most likely to recommend for gender affirming surgery out of the four they saw.

*Results*

We used a mixed effects regression to test whether gender identity and/or desire for gender-affirming surgery predicted participants' recommendations. We included a random intercept for participant, a fixed effect for gender identity (transgender woman [contrast-coded as 0.5] vs. non-binary person [contrast-coded as -0.5]), a fixed effect for target surgery desire (high desire [contrast-coded as 0.5] vs. regular desire [contrast-coded as -0.5]), and a fixed effect for the interaction of these two factors. We found that participants were overwhelmingly more likely to recommend those with high desire for surgery than those with regular desire for surgery ($b$ = 28.39, $SE$ = 1.29, $p < .001$) and also marginally more likely to recommend transgender women than nonbinary people ($b$ = 2.71, $SE$ = 1.29, $p$ = .04). However, recommendations did not differ significantly based on the interaction of gender and surgery desire ($p > .05$). Planned post-hoc analyses that looked at differences in surgery recommendation within each gender group found that both transgender and nonbinary people who had a high desire for surgery were more likely to get recommended for surgery ($t(594) = 16.03$, $p < .001$ for transgender people and $t(594)$ = 15.03, $p < .001$ for nonbinary people, respectively).

Additionally, when asked to select only one out of the four surgery seekers to recommend for surgery, we found that participants chose the transgender person with high desire for surgery over half the time, and much more than expected based on chance as indicated by a chi-square goodness-of-fit test($X^2$(3, N =199) = 162.65, $p$ < .001; also see Figure 5).

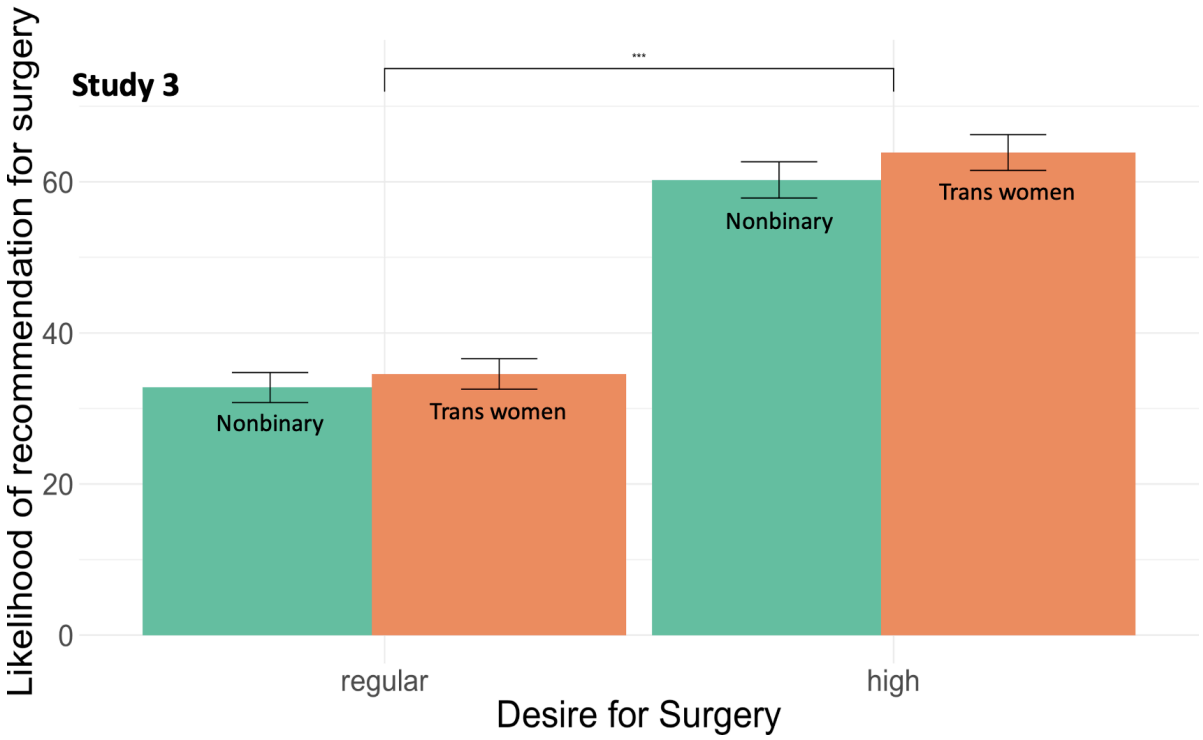



Figure 4. Likelihood of recommendation for surgery in Study 3. Recommendation for gender-affirming surgery was predicted by target's desire for surgery (high > regular), but not gender identity

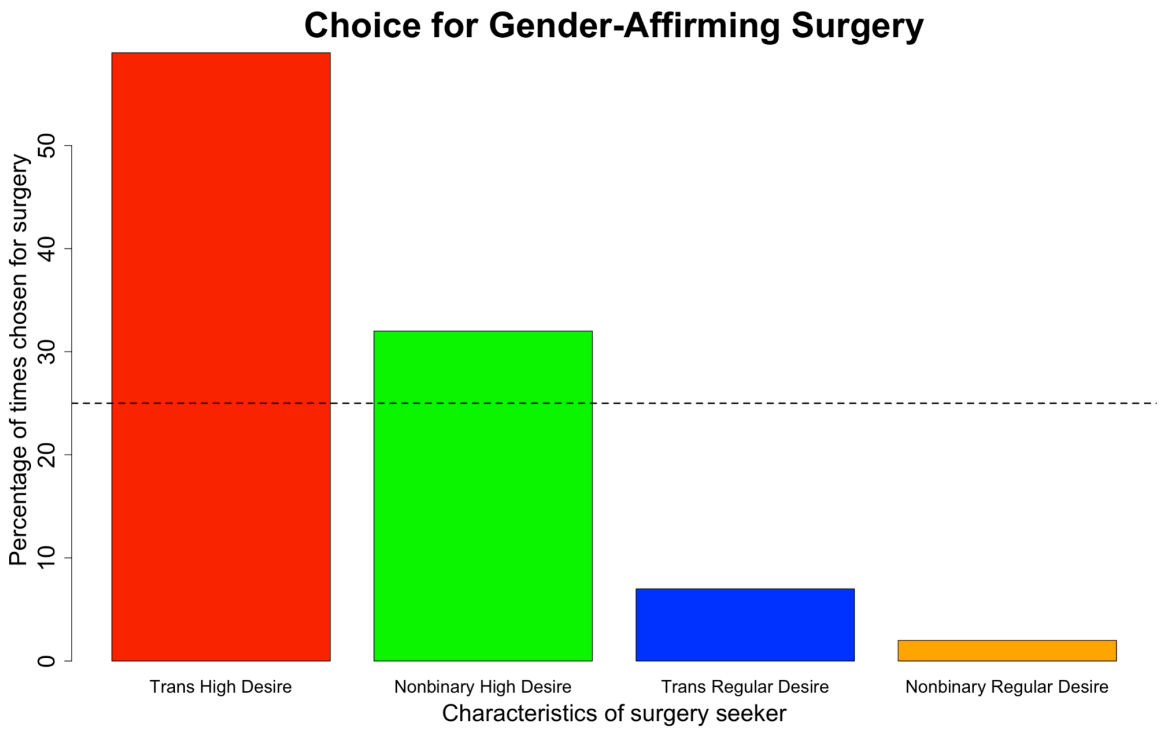



Figure 5. Percentage of participants selecting each type of person was selected when asked to choose only one out of four candidates for gender affirming surgery. Dotted line at 25% represents selections at chance.

*Discussion*

When making recommendations for gender affirming surgery, we find that cisgender participants overwhelmingly prioritize surgery seekers with a high desire for surgery over those with a regular desire for surgery. Additionally, we replicate our finding from Studies 1a-1c, finding again that cisgender participants prioritize binary trans people over nonbinary people, although the effect was less strong than in previous studies. Looking at these results in parallel, it appears that the effect of desire largely overshadows the effect of gender identity when prioritizing candidates for gender affirming surgery. However, when asked to choose a single person for gender affirming surgery, cisgender participants overwhelmingly chose a binary transgender person with a high desire for surgery over a nonbinary person who wanted surgery

just as much. This suggests there is still a strong tendency to favor binary transgender over nombinary surgical candidates.

## General Discussion

In sum, these studies suggest that in the absence of other information, cisgender people generally prioritize transgender women, individuals who first identified their gender in childhood or adolescence, and those whose appearance is stereotypically feminine for vaginoplasty than nonbinary people, individuals who first identified their gender recently, and those whose appearance is stereotypically masculine. This may be partially explained by cisgender people's motivation to prioritize those who they believe are more interested in surgery. This assumption may not always be right as there are, of course, nonbinary, later-identifying, and more masculine people who desire this surgery.

Further, our final study suggested that when matched for interest in surgery, participants were nearly twice as likely to recommend a transgender person than a nonbinary person. As a result, nonbinary surgery candidates who have a strong interest in surgery may have a more difficult time accessing gender affirming surgery, especially when the demand for surgery outpaces supply. Some research suggests that individuals seeking gender affirming care are aware of this bias and may respond by adjusting aspects of their identity and/or behavior in order to manage healthcare providers' impressions of them and maximize the likelihood of getting the healthcare they need, for example, by presenting as a more binary gender identity than they actually have (Wagner et al., 2016).

While our results suggest that cisgender participants' beliefs about who wants surgery may play some role in the observed results, they also suggest that there's something more at play given that binary and nonbinary targets who were matched for interest were evaluated differently.

Future work might examine how other possible factors such as the length of time a person has wanted surgery, a person's reason for wanting gender affirming surgery, and/or physical/mental health might intersect with gender identity and ultimately influence a person's likelihood of being recommended. For instance, because people tend to identify as nonbinary later in life (Tatum et al., 2020), surgery recommenders may automatically assume a trans person has waited longer for gender affirming surgery, and thus may prioritize trans surgery seekers over nonbinary surgery seekers. Additionally, a (binary) trans person may be more likely to want surgery to align their physical body with a binary gender (which aligns with common understandings of why people seek gender affirming surgeries), while a nonbinary person may seek gender affirming surgery for more diverse reasons such as a desire to explore their gender, become more androgynous, or achieve a genderless body (Galupo et al., 2021).

*Limitations*

The generality of our findings is inherently constrained by our participant sample - which consisted of primarily White cisgender heterosexual individuals on Prolific Academic and Cloud Research. On the one hand, White cisgender heterosexual people are a group who have more overall power in society, and are likely over-represented in the set of people who make medical decisions. However, the people on our sample were not specifically medical professionals, who arguably have the greatest decision-making power when it comes to the allocation of gender affirming surgery. While it's possible that their medical training or another unique feature of their role would discourage them from prioritizing, for example, binary over nonbinary patients, existing research on the paucity of education on gender-affirming healthcare in medical training would suggest that it's unlikely to be the case (Mousavian et al., 2024). If anything, experience with the medical model of gender affirming care may be associated with more gate-keeping of

individuals whose identity, expression, and/or behavior are less typical for these treatments (de Vries et al., 2020). Importantly, medical professionals are not the only people who play a role in the provision and care of trans and nonbinary people seeking surgery. Lay cisgender people may play other support and decision-making roles in regards to gender affirming care (e.g. as family members providing emotional and physical support, insurance claims adjusters, etc.).

Additionally, our studies were constrained by their design in that the surgery seekers evaluated were entirely fictitious, and the vignettes provided to participants contained significantly less information than would typically be available to a surgeon or healthcare professional. In real-world scenarios, the medical professionals responsible for making decisions about whether to allocate gender affirming care would typically have access to comprehensive medical histories as well as letters from mental health providers assessing their readiness for surgery (Coleman et al., 2022). Thus, it is possible that participants in our study were merely relying on their own stereotypes of the surgery seeker in the absence of other crucial information required to make decisions about providing surgery.

Finally, our design focused on a few aspects of candidates' backgrounds. There are many other aspects of their history, identity, or behavior that one could examine. For instance, in real life, surgery seekers are far more likely to be White than any other race, and young adults than older adults (Nolan et al., 2019; Lett et al., 2022). However, we did not manipulate surgery candidates' current age (all were young adults) or race (no information was provided) in our studies to examine if there are biases associated with these attributes. Future work might examine these and other biases that cisgender people might have in evaluating potential candidates for gender affirming surgery.

*Future Directions*

The majority of the data in the present set of studies suggests that participants were prioritizing surgery seekers' perceived needs by recommending those who they thought had the highest desire for surgery. However, in reality, there are many other factors that healthcare providers take into consideration when deciding whether or not to provide gender affirming care. Future work should examine how people factor in different (and sometimes conflicting) priorities when making decisions around gender affirming healthcare allocation. For example, how should a surgery seeker's motivation for seeking surgery be factored into the decision of whether or not to prioritize them for gender affirming care? For instance, should a person seeking gender affirming surgery to align their physical body with their internal sense of gender identity be automatically prioritized over a person seeking gender affirming surgery for the express purpose of gender exploration? What if the latter surgery candidate (focused on exploration) has reported greater desire for surgery than the former (focused on alignment)? What if the person exploration-focused candidate has severe chronic depression but also has a letter from their mental healthcare provider supporting their decision, while the alignment-focused surgery candidate has neither of these? In short, various aspects of a surgery candidate's identity and life can influence decisions about gender-affirming care, and future work should examine how candidates' desire for surgery interacts with these factors in the decision-making process.

Additionally, while the current study is focused on allocation of gender-affirming care of adults (18+), it would be interesting to see if people are guided by similar motivations when making decisions around (age-appropriate) gender affirming healthcare for adolescents (e.g. blockers, hormones, or surgery). Due to the current political discourse around gender-affirming healthcare for adolescents (Mahfouda et al., 2019; Kidd et al., 2021; de Vries et al., 2021; Hughes et al., 2021), it's likely that people would be overall more hesitant to recommend gender

affirming care to adolescents than adults. However, it would be interesting to determine whether people would still be more likely to recommend adolescents who have identified as binary trans from an early age and whose gender expression aligns with their current gender over those who have recently begun to identify as nonbinary and whose gender expression is more aligned with their assigned sex at birth for gender affirming medical care.

## Conclusion

In sum, these studies suggest that cisgender people prioritize binary transgender individuals, those who identify as transgender earlier, and those who have more feminine presentations for vaginoplasty than those who are nonbinary, identify as transgender later, and those with more masculine presentations. These effects may be partially driven by their perceptions of who wants surgery more. However, these assumptions may result in biases that do not apply to many of the people they impact, potentially putting some at a disadvantage when seeking gender affirming surgery.

## References


Agochukwu-Mmonu, N., Radix, A., Zhao, L., Makarov, D., Bluebond-Langner, R., Fendrick, A. M., Castle, E., & Berry, C. (2022). Patient reported outcomes in genital gender-affirming surgery: The time is now. *Journal of Patient-Reported Outcomes*, *6*(1). https://doi.org/10.1186/s41687-022-00446-x

Akhavan, A. A., Sandhu, S., Ndem, I., & Ogunleye, A. A. (2021). A review of gender affirmation surgery: What we know, and what we need to know. *Surgery*, *170*(1), 336–340. https://doi.org/10.1016/j.surg.2021.02.013

Almazan, A. N., & Keuroghlian, A. S. (2021). Association Between Gender-Affirming Surgeries and Mental Health Outcomes. *JAMA Surgery*, *156*(7), 611. https://doi.org/10.1001/jamasurg.2021.0952

Ashley, F. (2019). Gatekeeping hormone replacement therapy for transgender patients is dehumanising. *Journal of Medical Ethics*, *45*(7), 480–482. https://doi.org/10.1136/medethics-2018-105293

ASPS (American Society of Plastic Surgeons). (2017). *Plastic Surgery Statistics Report*. https://www.plasticsurgery.org/documents/News/Statistics/2017/plastic-surgery-statistics-full-report-2017.pdf

Bakko, M., & Kattari, S. K. (2020). Transgender-Related Insurance Denials as Barriers to Transgender Healthcare: Differences in Experience by Insurance Type. *Journal of General Internal Medicine*, *35*(6), 1693–1700. https://doi.org/10.1007/s11606-020-05724-2

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2024). *lme4: Linear Mixed-Effects Models using "Eigen" and S4* (Version 1.1-35.1) [Dataset]. https://doi.org/10.32614/CRAN.package.lme4

Berli, J. U., Knudson, G., Fraser, L., Tangpricha, V., Ettner, R., Ettner, F. M., Safer, J. D., Graham, J., Monstrey, S., & Schechter, L. (2017). What Surgeons Need to Know About Gender Confirmation Surgery When Providing Care for Transgender Individuals: A Review. *JAMA Surgery*, *152*(4), 394. https://doi.org/10.1001/jamasurg.2016.5549

Bettcher, T. M. (2014). Trapped in the Wrong Theory: Rethinking Trans Oppression and Resistance. *Signs: Journal of Women in Culture and Society*, *39*(2), 383–406. https://doi.org/10.1086/673088

Chu, J., Nagpal, M., & Dobberfuhl, A. D. (2024). Utilization and Cost of Gender-affirming Surgery in the United States from 2012-2019. *Annals of Surgery*. https://doi.org/10.1097/SLA.0000000000006296

Clark, B. A., Veale, J. F., Townsend, M., Frohard-Dourlent, H., & Saewyc, E. (2018). Non-binary youth: Access to gender-affirming primary health care. *International Journal of Transgenderism*, *19*(2), 158–169. https://doi.org/10.1080/15532739.2017.1394954

Cohen, W., Maisner, R. S., Mansukhani, P. A., & Keith, J. (2020). Barriers To Finding A Gender Affirming Surgeon. *Aesthetic Plastic Surgery*, *44*(6), 2300–2307. https://doi.org/10.1007/s00266-020-01883-z

Coleman, E., Radix, A. E., Bouman, W. P., Brown, G. R., de Vries, A. L. C., Deutsch, M. B., Ettner, R., Fraser, L., Goodman, M., Green, J., Hancock, A. B., Johnson, T. W., Karasic, D. H., Knudson, G. A., Leibowitz, S. F., Meyer-Bahlburg, H. F. L., Monstrey, S. J., Motmans, J., Nahata, L., … Arcelus, J. (2022). Standards of Care for the Health of Transgender and Gender Diverse People, Version 8. *International Journal of Transgender Health*, *23*(sup1), S1–S259. https://doi.org/10.1080/26895269.2022.2100644

De Vries, A. L. C., Richards, C., Tishelman, A. C., Motmans, J., Hannema, S. E., Green, J., & Rosenthal, S. M. (2021). Bell v Tavistock and Portman NHS Foundation Trust [2020] EWHC 3274: Weighing current knowledge and uncertainties in decisions about gender-related treatment for transgender adolescents. *International Journal of Transgender Health*, *22*(3), 217–224. https://doi.org/10.1080/26895269.2021.1904330

De Vries, E., Kathard, H., & Müller, A. (2020). Debate: Why should gender-affirming health care be included in health science curricula? *BMC Medical Education*, *20*(1). https://doi.org/10.1186/s12909-020-1963-6

dickey, lore m., Budge, S. L., Katz-Wise, S. L., & Garza, M. V. (2016). Health disparities in the transgender community: Exploring differences in insurance coverage. *Psychology of Sexual Orientation and Gender Diversity*, *3*(3), 275–282. https://doi.org/10.1037/sgd0000169

Downing, J., Holt, S. K., Cunetta, M., Gore, J. L., & Dy, G. W. (2022). Spending and Out-of-Pocket Costs for Genital Gender-Affirming Surgery in the US. *JAMA Surgery*, *157*(9), 799. https://doi.org/10.1001/jamasurg.2022.2606

Forcier, M. M., & Haddad, E. (2013). Health care for gender variant or gender non-conforming children. *Rhode Island Medical Journal (2013)*, *96*(4), 17–21.

Galupo, M. P., Cusack, C. E., & Morris, E. R. (2021). "Having a non-normative body for me is about survival": Androgynous body ideal among trans and nonbinary individuals. *Body Image*, *39*, 68–76. https://doi.org/10.1016/j.bodyim.2021.06.003

Glover, J. K. (2016). Redefining Realness?: On Janet Mock, Laverne Cox, TS Madison, and the Representation of Transgender Women of Color in Media. *Souls*, *18*(2–4), 338–357. https://doi.org/10.1080/10999949.2016.1230824

Houssayni, S., & Nilsen, K. (2018). Transgender Competent Provider: Identifying Transgender Health Needs, Health Disparities, and Health Coverage. *Kansas Journal of Medicine*, *11*(1), 1–18.

Hughes, L. D., Kidd, K. M., Gamarel, K. E., Operario, D., & Dowshen, N. (2021a). "These Laws Will Be Devastating": Provider Perspectives on Legislation Banning Gender-Affirming Care for Transgender Adolescents. *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, *69*(6), 976–982. https://doi.org/10.1016/j.jadohealth.2021.08.020

Hughes, L. D., Kidd, K. M., Gamarel, K. E., Operario, D., & Dowshen, N. (2021b). "These Laws Will Be Devastating": Provider Perspectives on Legislation Banning Gender-Affirming Care for Transgender Adolescents. *Journal of Adolescent Health*, *69*(6), 976–982. https://doi.org/10.1016/j.jadohealth.2021.08.020

Johnson, A. H. (2015). Normative Accountability: How the Medical Model Influences Transgender Identities and Experiences: Normative Accountability. *Sociology Compass*, *9*(9), 803–813. https://doi.org/10.1111/soc4.12297

Kattari, S. K., Atteberry-Ash, B., Kinney, M. K., Walls, N. E., & Kattari, L. (2019). One size does not fit all: Differential transgender health experiences. *Social Work in Health Care*, *58*(9), 899–917. https://doi.org/10.1080/00981389.2019.1677279

Kattari, S. K., & Hasche, L. (2016). Differences Across Age Groups in Transgender and Gender Non-Conforming People's Experiences of Health Care Discrimination, Harassment, and Victimization. *Journal of Aging and Health*, *28*(2), 285–306. https://doi.org/10.1177/0898264315590228

Kidd, K. M., Sequeira, G. M., Paglisotti, T., Katz-Wise, S. L., Kazmerski, T. M., Hillier, A., Miller, E., & Dowshen, N. (2021). “This Could Mean Death for My Child”: Parent Perspectives on Laws Banning Gender-Affirming Care for Transgender Adolescents. *Journal of Adolescent Health*, *68*(6), 1082–1088. https://doi.org/10.1016/j.jadohealth.2020.09.010

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2020). *lmerTest: Tests in Linear Mixed Effects Models* (Version 3.1-3) [Dataset]. https://doi.org/10.32614/CRAN.package.lmerTest

Lefevor, G. T., Boyd-Rogers, C. C., Sprague, B. M., & Janis, R. A. (2019). Health disparities between genderqueer, transgender, and cisgender individuals: An extension of minority stress theory. *Journal of Counseling Psychology*, *66*(4), 385–395. https://doi.org/10.1037/cou0000339

Lenth, R. V. (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (Version 1.10.1) [Dataset]. https://doi.org/10.32614/CRAN.package.emmeans

Lerner, J. E., & Robles, G. (2017). Perceived Barriers and Facilitators to Health Care Utilization in the United States for Transgender People: A Review of Recent Literature. *Journal of Health Care for the Poor and Underserved*, *28*(1), 127–152. https://doi.org/10.1353/hpu.2017.0014

Lett, E., Abrams, M. P., Gold, A., Fullerton, F.-A., & Everhart, A. (2022). Ethnoracial inequities in access to gender-affirming mental health care and psychological distress among transgender adults. *Social Psychiatry and Psychiatric Epidemiology*, *57*(5), 963–971. https://doi.org/10.1007/s00127-022-02246-6

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

Loo, S., Almazan, A. N., Vedilago, V., Stott, B., Reisner, S. L., & Keuroghlian, A. S. (2021). Understanding community member and health care professional perspectives on gender-affirming care—A qualitative study. *PLOS ONE*, *16*(8), e0255568. https://doi.org/10.1371/journal.pone.0255568

Lüdecke, D. (2023). *sjPlot: Data Visualization for Statistics in Social Science* (Version 2.8.15) [Dataset]. https://doi.org/10.32614/CRAN.package.sjPlot

Mahfouda, S., Moore, J. K., Siafarikas, A., Hewitt, T., Ganti, U., Lin, A., & Zepf, F. D. (2019). Gender-affirming hormones and surgery in transgender children and adolescents. *The Lancet Diabetes & Endocrinology*, *7*(6), 484–498. https://doi.org/10.1016/S2213-8587(18)30305-X

Morgenroth, T., Kirby, T. A., Gee, I. A., & Ovett, T. A. (2021). Born this Way–or Not? The Relationship Between Essentialism and Sexual Minorities' LGBTQ+ Identification and Belonging. *Archives of Sexual Behavior*, *50*(8), 3447–3458. https://doi.org/10.1007/s10508-021-02145-y

Mousavian, M., Ranganathan, K., Keuroghlian, A. S., Park, Y. S., & Kumar, A. (2024). What are the barriers to health professionals' training on gender-affirming care from patients' and clinicians' perspectives? *Social Science & Medicine*, *351*, 116983. https://doi.org/10.1016/j.socscimed.2024.116983

Nguyen, H. B., Chavez, A. M., Lipner, E., Hantsoo, L., Kornfield, S. L., Davies, R. D., & Epperson, C. N. (2018). Gender-Affirming Hormone Use in Transgender Individuals:

Impact on Behavioral Health and Cognition. *Current Psychiatry Reports*, *20*(12), 110. https://doi.org/10.1007/s11920-018-0973-0

Nolan, I. T., Dy, G. W., & Levitt, N. (2019). Considerations in Gender-Affirming Surgery. *Urologic Clinics of North America*, *46*(4), 459–465. https://doi.org/10.1016/j.ucl.2019.07.004

Nolan, I. T., Kuhner, C. J., & Dy, G. W. (2019). Demographic and temporal trends in transgender identities and gender confirming surgery. *Translational Andrology and Urology*, *8*(3), 184–190. https://doi.org/10.21037/tau.2019.04.09

Puckett, J. A., Cleary, P., Rossman, K., Mustanski, B., & Newcomb, M. E. (2018). Barriers to Gender-Affirming Care for Transgender and Gender Nonconforming Individuals. *Sexuality Research and Social Policy*, *15*(1), 48–59. https://doi.org/10.1007/s13178-017-0295-8

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research* (Version 2.3.9) [Computer software]. https://personality-project.org/r/psych/ https://personality-project.org/r/psych-manual.pdf

Rich, B. (2023). *table1: Tables of Descriptive Statistics in HTML* (Version 1.4.3) [Computer software]. https://rdocumentation.org/packages/table1/versions/1.4.3

Roberts, S. O., Ho, A. K., Rhodes, M., & Gelman, S. A. (2017). Making Boundaries Great Again: Essentialism and Support for Boundary-Enhancing Initiatives. *Personality and Social Psychology Bulletin*, *43*(12), 1643–1658. https://doi.org/10.1177/0146167217724801

RStudio Team. (2023). *RStudio: Integrated development environment for R* (Version 2023.09.1+494) [Computer software]. Posit Software, PBC. https://posit.co/download/rstudio/

Safer, J. D., Coleman, E., Feldman, J., Garofalo, R., Hembree, W., Radix, A., & Sevelius, J. (2016). Barriers to healthcare for transgender individuals: *Current Opinion in Endocrinology & Diabetes and Obesity*, *23*(2), 168–171. https://doi.org/10.1097/MED.0000000000000227

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining Power and Sample Size for Simple and Complex Mediation Models. *Social Psychological and Personality Science*, *8*(4), 379–386. https://doi.org/10.1177/1948550617715068

Sevelius, J., & Jenness, V. (2017). Challenges and opportunities for gender-affirming healthcare for transgender women in prison. *International Journal of Prisoner Health*, *13*(1), 32–40. https://doi.org/10.1108/IJPH-08-2016-0046

Steensma, T. D., McGuire, J. K., Kreukels, B. P. C., Beekman, A. J., & Cohen-Kettenis, P. T. (2013). Factors Associated With Desistence and Persistence of Childhood Gender Dysphoria: A Quantitative Follow-Up Study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(6), 582–590. https://doi.org/10.1016/j.jaac.2013.03.016

Swan, J., Phillips, T. M., Sanders, T., Mullens, A. B., Debattista, J., & Brömdal, A. (2022). Mental health and quality of life outcomes of gender-affirming surgery: A systematic literature review. *Journal of Gay & Lesbian Mental Health*, 1–44. https://doi.org/10.1080/19359705.2021.2016537

Tatum, A. K., Catalpa, J., Bradford, N. J., Kovic, A., & Berg, D. R. (2020). Examining identity development and transition differences among binary transgender and genderqueer

nonbinary (GQNB) individuals. *Psychology of Sexual Orientation and Gender Diversity*, *7*(4), 379–385. https://doi.org/10.1037/sgd0000377

Tee, N., & Hegarty, P. (2006). Predicting opposition to the civil rights of trans persons in the United Kingdom. *Journal of Community & Applied Social Psychology*, *16*(1), 70–80. https://doi.org/10.1002/casp.851

Timothy Cavanaugh, Ruben Hopwood, & Cei Lambert. (2016). Informed Consent in the Medical Care of Transgender and Gender-Nonconforming Patients. *AMA Journal of Ethics*, *18*(11), 1147–1155. https://doi.org/10.1001/journalofethics.2016.18.11.sect1-1611

van den Berg, M., & Marinus, M. (2017). Trans scripts: The representation of transgender people in the media in the Netherlands (1991-2016). *Tijdschrift Voor Genderstudies*, *20*(4), 379–397. https://doi.org/10.5117/TVGN2017.4.BERG

Wagner, P. E., Kunkel, A., Asbury, M. B., & Soto, F. (2016). Health (trans) gressions: Identity and stigma management in trans* healthcare support seeking. *Women & Language*, *39*(1), 49–74.

Weber, U., Mummendey, A., & Waldzus, S. (2002). Perceived legitimacy of intergroup status differences: Its prediction by relative ingroup prototypicality. *European Journal of Social Psychology*, *32*(4), 449–470. https://doi.org/10.1002/ejsp.102

Wernick, J. A., Busa, S., Matouk, K., Nicholson, J., & Janssen, A. (2019). A Systematic Review of the Psychological Benefits of Gender-Affirming Surgery. *Urologic Clinics of North America*, *46*(4), 475–486. https://doi.org/10.1016/j.ucl.2019.07.002

Wickham, H. (2023). *tidyverse: Easily Install and Load the "Tidyverse"* (Version 2.0.0) [Dataset]. https://doi.org/10.32614/CRAN.package.tidyverse

Wilton, L. S., Bell, A. N., Carpinella, C. M., Young, D. M., Meyers, C., & Clapham, R. (2019). Lay Theories of Gender Influence Support for Women and Transgender People's Legal Rights. *Social Psychological and Personality Science*, *10*(7), 883–894. https://doi.org/10.1177/1948550618803608

Xie, Y. (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R* (Version 1.45) [Dataset]. https://doi.org/10.32614/CRAN.package.knitr

**Supplementary Material for *Cisgender people's perceptions of candidates for gender affirming surgery***

Table of Contents

**Additional Clarifications about Analytic Approach..... 227**
**Additional Information about Studies 1a..... 227**
Study 1a Preregistered Analyses..... 227
Study 1a Additional Preregistered Analyses..... 228
Study 1a Full Output of All Mixed Effects Models Reported in Main Manuscript..... 228
**Additional Information about Study 1b..... 230**
Study 1b Preregistered Analyses..... 230
Study 1b Full Output of All Mixed Effects Models Reported in Main Manuscript..... 231
**Additional Information about Study 1c..... 233**
Study 1c Detailed Procedure..... 233
Study 1c Additional Measures..... 233
Study 1c Preregistered Analyses..... 234
Study 1c Additional Preregistered Analyses..... 235
Study 1c Full Output of All Mixed Effects Models Reported in Main Manuscript..... 236
**Additional Information about Study 2..... 237**
Study 2 Detailed Method..... 237
Study 2 Full Analyses..... 237
**Additional Information about Study 3..... 240**
Study 3 Example Surgery Readiness Forms..... 240
Study 3 Preregistered Analyses..... 245
Study 3 Full Output of All Mixed Effects Models Reported in Main Manuscript..... 245

## Additional Clarifications About Analytic Approach

Please note that, in our preregistrations for studies 1a and 1b, we referred to both gender expression (feminine vs. masculine) and age of identification (in childhood vs. recently) as 'prototypicality'. Each of the eight trials is defined with respect to only one or the other; these two versions of prototypicality and/or expected prioritization are not crossed. Additionally, all analyzed trials were blocked according to this definition of prototypicality – a participant either saw all prototypical trials first (regardless of vignette gender identity) or all non-prototypical trials first (regardless of gender identity).

Though we did not define how we would code our predictors in the preregistrations, we did say we would consider our hypotheses with respect to main effects. We therefore contrast code our predictors in studies 1a, 1b, 1c, and 3 (as described in the manuscript) so that we can interpret our regression coefficients as main (not conditional) effects.

Finally, we preregistered post-hoc t-tests in Studies 1a, 1b, and 3 following our mixed effects models. We run these as independent samples t-tests in the supplement to maintain highest fidelity to the original preregistration document. However, we acknowledge that running independent samples t-tests on data with multiple within-participant observations yields uninterpretable statistics. Thus, we refer readers to the model-based contrasts presented in the main paper for more reliable insight on these comparisons.

## Additional Information about Study 1a

### *Preregistered Analyses*

We will run the following mixed effects model on two different subsets of data. First we will run it looking at all prototypicality manipulation data, and second, looking at only prototypicality manipulation data within the first block. In both instances, we will predict likelihood of surgery recommendation from target gender, prototypicality level, and their interaction. To do this, we will use a mixed effects model (listed below) which will include fixed effects of target gender, prototypicality, and target gender*prototypicality. The model will also include a random intercept for subject id.

*surgery_recommendation ~ target_gender*prototypicality_level + (1| subject_id)*

The results of the first mixed effects model using all data revealed that participants were more likely to recommend surgery candidates with prototypical characteristics than those with non-prototypical characteristics, ($b$ = 13.29, $SE$ = 1.05, $p$ < .001), and more likely to recommend transgender women than nonbinary people ($b$ = 13.36, $SE$ = 1.05, $p$ < .001). Finally, there is an interaction between prototypicality and target gender such that the extent to which participants were more likely to recommend prototypical over non-prototypical surgery candidates was greater for those who were transgender than for those who were nonbinary; $b$ = 8.50, $SE$ = 2.09, $p$ < .001).

The results of the second mixed effects model using data only within the first block revealed participants were more likely to recommend surgery candidates with prototypical than non-prototypical characteristics, ($b$ = 8.25, $SE$ = 3.96, $p$ = .038), and more likely to recommend transgender women than nonbinary people ($b$ = 13.97, $SE$ = 1.47, $p$ < .001). Finally, there is an interaction between prototypicality and target gender such that the extent to which participants were more likely to recommend prototypical over non-prototypical surgery candidates was greater for those who were transgender than for those who were nonbinary; $b$ = 12.68, $SE$ = 2.95, $p$ < .001).

We will do 2 post-hoc t-tests[27] comparing whether or not likelihood of recommendation differs for either trans people or for nonbinary people characterised as prototypical vs. non-prototypical.

For trans surgery seekers, participants were more likely to recommend people characterised as prototypical over those characterised as non–prototypical, $t(811.36) = 7.44$, $p < .001$. Similarly, for nonbinary surgery seekers, participants were more likely to recommend people characterised as prototypical over those characterised as non–prototypical, $t(813.8) = 3.85$, $p < .001$.

For any of the aforementioned post-hoc t-tests that are significant, we will do an additional two t-tests[1] comparing likelihood of recommendation for the high vs. low prototypicality separately for age of first identification and gender expression.

Age of first identification: For trans and nonbinary surgery seekers, participants were more likely to recommend those who first recognized their gender identity in childhood over those who first recognized their gender identity recently, (transgender women: $t(405.83) = 8.53$, $p < .001$; nonbinary: $t(403.57) = 3.76$, $p < .001$).

Gender expression: For trans surgery seekers, participants were more likely to recommend those characterised as stereotypically feminine over those characterised as stereotypically masculine, $t(405.68) = 2.23$, $p = .026$. Conversely, for nonbinary surgery seekers, participants were no more likely to recommend those characterised as stereotypically feminine than those characterised as stereotypically masculine, $t(405.99) = 1.85$, $p = .065$.

***Additional Preregistered Analyses***

In order to determine whether the order in which statements are shown impacts people's responses, we will run an additional model on the prototypicality manipulation data, predicting likelihood of surgery recommendation from order, prototypicality level, and their interaction. To do this, we will use a mixed effects model (listed below) which will include fixed effects of order, prototypicality, and order*prototypicality. As before, the model will also include a random intercept for subject id.

*surgery_recommendation ~ order*prototypicality_level + (1| subject_id)*

The results of this mixed effects model revealed participants show no difference in likelihood of recommending surgery regardless of whether they viewed prototypical or non-prototypical candidates first, ($b = -5.04$, $SE = 3.80$, $p = .185$). Additionally, participants were more likely to recommend surgery candidates with prototypical characteristics than those with non-prototypical characteristics, ($b = 13.29$, $SE = 1.11$, $p < .001$).

***Full Output of All Mixed Effects Models Reported in Main Manuscript***

Model 1: Mixed effects model: gender identity x prototypicality (preregistered)

This mixed effects regression included a random intercept for participant and three fixed effect predictors: gender identity (transgender woman [contrast-coded as 0.5] vs. non-binary person [contrast-coded as -0.5]), non-gender-identity expected prioritization (feminine expression or early identification [contrast-coded as 0.5] vs. masculine expression or recent identification [contrast-coded as -0.5]), and the interaction between these two factors.

**surgery_recommendation**

[27] See note about post-hoc t-tests in the final paragraph of Additional Details About Analytic Approach above

| *Predictors* | *Estimates* | *std. Error* | *CI* | *p* |
|---|---|---|---|---|
| (Intercept) | 54.57 | 1.91 | 50.84 – 58.31 | **<0.001** |
| prototypicality level | 13.29 | 1.05 | 11.24 – 15.35 | **<0.001** |
| target gender | 13.36 | 1.05 | 11.30 – 15.41 | **<0.001** |
| prototypicality level × target gender | 8.50 | 2.09 | 4.39 – 12.61 | **<0.001** |
| **Random Effects** | | | | |
| σ2 | 447.67 | | | |
| τ00 workerId | 684.62 | | | |
| ICC | 0.60 | | | |
| N workerId | 204 | | | |
| Observations | 1632 | | | |
| Marginal R2 / Conditional R2 | 0.076 / 0.635 | | | |

Model 2: Mixed effects model: gender expression (not preregistered)
This mixed effects regression included a random intercept for participant and three fixed effect predictors: gender identity (transgender woman [contrast-coded as 0.5] vs. non-binary person [contrast-coded as -0.5]), gender expression (feminine expression [contrast-coded as 0.5] vs. masculine expression [contrast-coded as -0.5]), and the interaction between these two factors.

| | **surgery_recommendation** | | | |
|---|---|---|---|---|
| *Predictors* | *Estimates* | *std. Error* | *CI* | *p* |
| (Intercept) | 60.90 | 2.03 | 56.92 – 64.87 | **<0.001** |
| prototypicality level | 6.69 | 1.31 | 4.11 – 9.27 | **<0.001** |
| target gender | 12.46 | 1.31 | 9.88 – 15.04 | **<0.001** |
| prototypicality level × target gender | 1.02 | 2.63 | -4.13 – 6.18 | 0.697 |
| **Random Effects** | | | | |
| σ2 | 352.22 | | | |
| τ00 workerId | 749.67 | | | |
| ICC | 0.68 | | | |
| N workerId | 204 | | | |
| Observations | 816 | | | |

| | |
|---|---|
| Marginal R2 / Conditional R2 | 0.044 / 0.694 |

Model 3: Mixed effects model: age of identification (not preregistered)
This mixed effects regression included a random intercept for participant and three fixed effect predictors: gender identity (transgender woman [contrast-coded as 0.5] vs. non-binary person [contrast-coded as -0.5]), first gender identification (early identification [contrast-coded as 0.5] vs. recent identification [contrast-coded as -0.5]), and the interaction between these two factors.

| | surgery_recommendation | | | |
|---|---|---|---|---|
| *Predictors* | *Estimates* | *std. Error* | *CI* | *p* |
| (Intercept) | 48.25 | 1.92 | 44.49 – 52.02 | **<0.001** |
| prototypicality level | 19.89 | 1.41 | 17.12 – 22.66 | **<0.001** |
| target gender | 14.25 | 1.41 | 11.48 – 17.03 | **<0.001** |
| prototypicality level × target gender | 15.97 | 2.82 | 10.43 – 21.51 | **<0.001** |
| **Random Effects** | | | | |
| σ2 | 406.37 | | | |
| τ00 workerId | 648.89 | | | |
| ICC | 0.61 | | | |
| N workerId | 204 | | | |
| Observations | 816 | | | |
| Marginal R2 / Conditional R2 | 0.136 / 0.667 | | | |

**Additional Information about Study 1b**

***Preregistered Analyses***

We will run the following mixed effects model on two different subsets of data. First we will run it looking at all data, and second, looking at only data from the first block. In both instances, we will replicate the analysis used in our previous study and predict likelihood of surgery recommendation from target gender, prototypicality level, and their interaction. To do this, we will use a mixed effects model (listed below) which will include fixed effects of target gender, prototypicality, and target gender*prototypicality. The model will also include a random intercept for subject id.

*surgery_recommendation ~ target_gender*prototypicality_level + (1| subject_id)*

The results of the first mixed effects model using all data revealed that like in Study 1a, participants in Study 1b are more likely to recommend surgery candidates with prototypical characteristics than those with non-prototypical characteristics, ($b$ = 15.17, $SE$ = 1.09, $p$ < .001), and more likely to recommend surgery candidates who are (binary) transgender than those who are nonbinary ($b$ = 12.99, $SE$ = 1.09, $p$ < .001). Finally, there is interaction between prototypicality and target gender such

that the extent to which participants were more likely to recommend prototypical over non-prototypical surgery candidates was greater for those who were transgender than for those who were nonbinary; $b$ = 4.63, $SE$ = 2.18, $p$ = .034).

The results of the second mixed effects model using data only within the first block revealed participants were more likely to recommend surgery candidates with prototypical than non-prototypical characteristics, ($b$ = 10.18, $SE$ = 3.65, $p$ = .005), and more likely to recommend transgender women than nonbinary people ($b$ = 15.07, $SE$ = 1.46, $p$ < .001). Finally, there is an interaction between prototypicality and target gender such that the extent to which participants were more likely to recommend prototypical over non-prototypical surgery candidates was greater for those who were transgender than for those who were nonbinary; $b$ = 13.24, $SE$ = 2.92, $p$ < .001).

We will also do 2 post-hoc t-tests[1] comparing whether or not likelihood of recommendation differs for either trans people or for nonbinary people characterised as high vs. low prototypicality.

Like in Study 1a, participants in Study 1b were more likely to recommend trans surgery seekers characterised as prototypical over those characterised as non–prototypical, $t$(803.26) = 7.74, $p$ < .001. Similarly, for nonbinary surgery seekers, participants were more likely to recommend people characterised as prototypical over those characterised as non–prototypical, $t$(805.4) = 5.80, $p$ < .001.

For any of the aforementioned post-hoc t-tests that are significant, we will do an additional two t-tests[1] comparing likelihood of recommendation for the high vs. low prototypicality separately for age of first identification and gender expression.

Age of first identification: For trans and nonbinary surgery seekers, participants were more likely to recommend those who first recognized their gender identity in childhood over those who first recognized their gender identity recently, ($t$(401.6) = 9.01, $p$ < .001 for trans surgery seekers and $t$(399.19) = 5.48, $p$ < .001 for nonbinary surgery seekers).

Gender expression: For both trans and nonbinary surgery seekers, participants were more likely to recommend those characterised as stereotypically feminine over those characterised as stereotypically masculine ($t$(401.95) = 2.32, $p$ = .02 for trans surgery seekers and $t$(401.96) = 2.85, $p$ = .004 for nonbinary surgery seekers).

***Full Output of All Mixed Effects Models Reported in Main Manuscript***

Model 1: Mixed effects model: gender identity x prototypicality (preregistered)

This mixed effects regression included a random intercept for participant and three fixed effect predictors: gender identity (transgender woman [contrast-coded as 0.5] vs. non-binary person [contrast-coded as -0.5]), non-gender-identity expected prioritization (feminine expression or early identification [contrast-coded as 0.5] vs. masculine expression or recent identification [contrast-coded as -0.5]), and the interaction between these two factors.

| | surgery_recommendation | | | |
|---|---|---|---|---|
| *Predictors* | *Estimates* | *std. Error* | *CI* | *p* |
| (Intercept) | 51.95 | 1.72 | 48.59 – 55.32 | **<0.001** |
| prototypicality level | 15.17 | 1.09 | 13.03 – 17.31 | **<0.001** |

| | | | | |
|---|---|---|---|---|
| target gender | 12.99 | 1.09 | 10.85 – 15.12 | **<0.001** |
| prototypicality level × target gender | 4.63 | 2.18 | 0.35 – 8.90 | **0.034** |

**Random Effects**

| | |
|---|---|
| σ2 | 479.60 |
| τ00 workerId | 534.53 |
| ICC | 0.53 |
| N workerId | 202 |
| Observations | 1616 |
| Marginal R2 / Conditional R2 | 0.091 / 0.570 |

Model 2: Mixed effects model: gender expression (not preregistered)
This mixed effects regression included a random intercept for participant and three fixed effect predictors: gender identity (transgender woman [contrast-coded as 0.5] vs. non-binary person [contrast-coded as -0.5]), gender expression (feminine expression [contrast-coded as 0.5] vs. masculine expression [contrast-coded as -0.5]), and the interaction between these two factors.

| | **surgery_recommendation** | | | |
|---|---|---|---|---|
| *Predictors* | *Estimates* | *std. Error* | *CI* | *p* |
| (Intercept) | 56.70 | 1.83 | 53.10 – 60.30 | **<0.001** |
| prototypicality level | 8.15 | 1.47 | 5.26 – 11.04 | **<0.001** |
| target gender | 13.60 | 1.47 | 10.71 – 16.49 | **<0.001** |
| prototypicality level × target gender | -1.60 | 2.95 | -7.38 – 4.19 | 0.588 |

**Random Effects**

| | |
|---|---|
| σ2 | 438.63 |
| τ00 workerId | 569.78 |
| ICC | 0.57 |
| N workerId | 202 |
| Observations | 808 |
| Marginal R2 / Conditional R2 | 0.059 / 0.591 |

Model 3: Mixed effects model: first identification (not preregistered)
This mixed effects regression included a random intercept for participant and three fixed effect predictors: gender identity (transgender woman [contrast-coded as 0.5] vs. non-binary person [contrast-coded as

-0.5]), first gender identification (early identification [contrast-coded as 0.5] vs. recent identification [contrast-coded as -0.5]), and the interaction between these two factors.

| | surgery_recommendation | | | |
|---|---|---|---|---|
| *Predictors* | *Estimates* | *std. Error* | *CI* | *p* |
| (Intercept) | 47.21 | 1.72 | 43.84 – 50.59 | **<0.001** |
| prototypicality level | 22.19 | 1.52 | 19.21 – 25.17 | **<0.001** |
| target gender | 12.37 | 1.52 | 9.39 – 15.35 | **<0.001** |
| prototypicality level × target gender | 10.85 | 3.04 | 4.89 – 16.81 | **<0.001** |
| **Random Effects** | | | | |
| σ2 | 465.68 | | | |
| τ00 workerId | 481.52 | | | |
| ICC | 0.51 | | | |
| N workerId | 202 | | | |
| Observations | 808 | | | |
| Marginal R2 / Conditional R2 | 0.151 / 0.583 | | | |

**Additional Information about Study 1c**

***Detailed Procedure***

As in Studies 1a and 1b, we asked participants in Study 1c to read vignettes about trans and nonbinary people and make recommendations for gender affirming surgery. We manipulated the age at which the individuals in the vignettes first realized their gender to be either during early childhood (in alignment with essentialist reasoning) or around adolescence (in alignment with pilot data from 51 individuals suggesting that the average age at which trans and nonbinary individuals are expected to have first realized their identity was 14.56 and 13.64 years, respectively[28]). Then, we measured participants' endorsement of essentialism and prototypicality in relation to the gender of the individuals in the vignettes and used a parallel mediation model to test whether either endorsement of essentialism or prototypicality mediated the relationship between age of first identification and likelihood of surgery recommendation.

***Additional Measures***

[28] In this short pilot study, participants (N = 51) were asked to imagine either an "average transgender woman" or "average nonbinary person assigned male at birth" (e.g. not a celebrity or somebody they knew personally) and then share the approximate current age of the person they imagined as well as the approximate age of this person (in years) when they first began to feel they were a girl (or nonbinary). Overall, participants imagined the average trans woman as a current 28.8 year-old (SD 6.29 years) who first felt like a girl around age 14.56 (SD 6.08 years), and the average nonbinary person assigned male at birth as a current 25 year old (SD 4.43 years) who first felt like not a boy or girl around age 13.64 (SD 3.98 years).

In Study 1c, participants read a vignette about either a trans person or a nonbinary person assigned male at birth interested in having gender affirming surgery and assessed whether or not the individual in the vignette should be recommended for surgery. Additionally, all participants completed the following questions[29] designed to tap into their beliefs around essentialism (labeled E below) or prototypicality (labeled P below) based on their opinion of the individual they just recommended for surgery on a 6 point scale from *This is definitely not true* to *This is definitely true:*

- It's impossible for [individual] to truly change her gender identity (E)
- [Individual] chose her gender identity (E, reverse-coded)
- Just knowing [individual's] gender identity makes it easy to predict many aspects of her personality (E)
- Even if you know [individual's] gender identity, it is never easy to make accurate judgments about how she will act in different situations. (E, reverse-coded)
- [Individual] was probably born the way she is in terms of her gender identity. (E)
- There's a biological reason why [individual] has her gender identity. (E)
- If [individual] had had a different set of life experiences, she would probably have a different gender identity. (E, reverse coded)
- [Individual] is a pretty typical trans woman. (P)
- [Individual] is quite different than most trans women. (P, reverse coded)
- [Individual's] life experiences seem pretty similar to those of most trans women. (P)

***Preregistered Analyses***

We will run the following mixed effects model* (listed below) which will include fixed effects of target gender (trans vs. nonbinary), age of first identification (childhood vs. adolescence), and target gender*age of first identification. The model will also include a random intercept for subject id.

*surgery_recommendation ~ target_gender * age_of_first_identification + (1| subject_id)*

*Note that we did *not* run our pre-registered model because our dataset did not include repeated measures. Instead we ran the following linear model that did not include any random effects:

*surgery_recommendation ~ target_gender * age_of_first_identification*

The results of this linear model revealed that, like before, participants in Study 1c are more likely to recommend surgery candidates who are (binary) transgender than those who are nonbinary ($b$ = 12.17, $SE$ = 2.55, $p$ < .001). However, participants showed no difference in likelihood of surgery recommendation for surgery seekers who first came out in childhood vs. adolescence ($b$ = -1.10, $SE$ = 2.55, $p$ = .666). Finally, there was no interaction between age of first gender identification and target gender; $b$ = 2.66, $SE$ = 5.11 $p$ = .602).

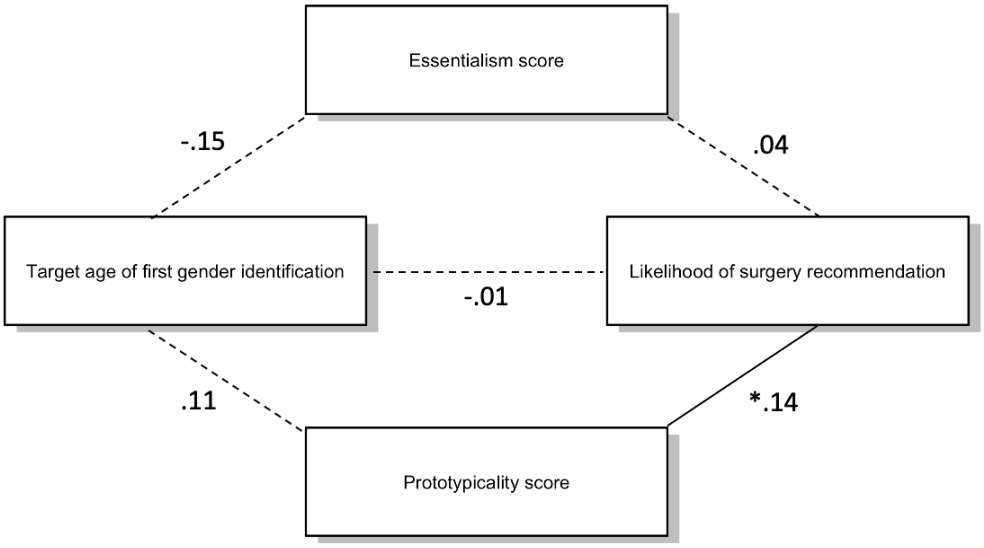
We will run the following mediation model separately on data from participants in the trans vs. nonbinary conditions:

We will use a PROCESS macro in R to test whether likelihood of surgery recommendation differs between different ages of first identification (childhood vs. adolescent). In this model, we will enter age of first identification as the predictor, essentialism score (participants' average essentialism ratings) and prototypicality score (participants' average prototypicality ratings) as parallel mediators, and likelihood of surgery recommendation as the outcome.

[29] Questions were customized to use she/her and "trans woman" when asked to in reference to trans individuals and use they/them and "nonbinary person" when asked in reference to nonbinary individuals

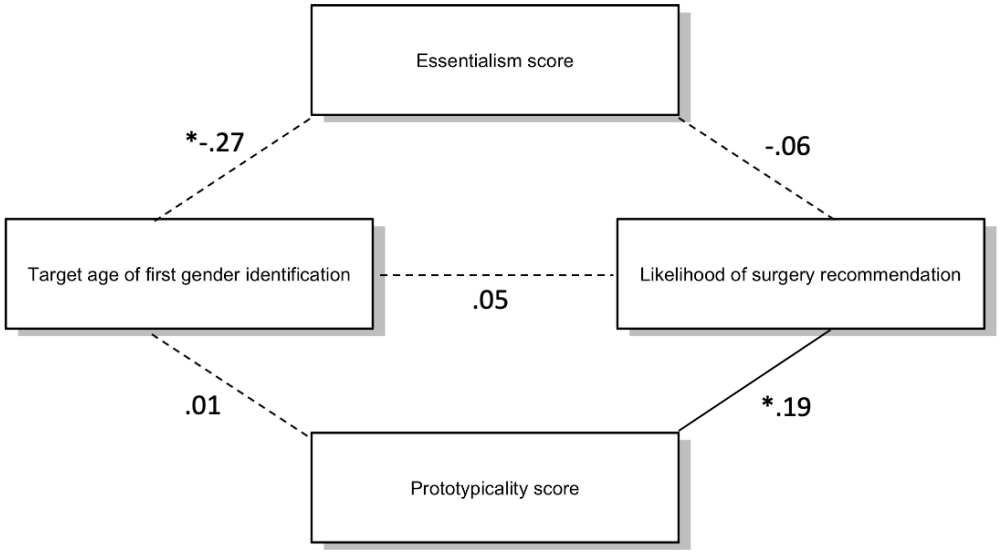
Transgender Targets: We found no evidence of an indirect effect of age of first identification on surgery recommendation through either essentialism or prototypicality ratings (see Figure S1 below).

Figure S1. Parallel Mediation with data from transgender targets




Nonbinary Targets: We found no evidence of an indirect effect of age of first identification on surgery recommendation through either essentialism or prototypicality ratings (see Figure S2 below).

Figure S2. Parallel Mediation with data from nonbinary targets




***Additional Preregistered Analyses***

Before running Analysis 2, we will run three correlations: one between ratings of essentialism and prototypicality, one between ratings of essentialism and likelihood of recommendation, and one between ratings of prototypicality and likelihood of recommendation.

- There was a (weakly) positive correlation between ratings of essentialism and prototypicality, $r_S$ = .08, $p$ = .041.
- There was no correlation between ratings of essentialism and likelihood of surgery recommendation, $r_S$ = .01, $p$ = .750.
- There was a positive correlation between ratings of prototypicality and likelihood of surgery recommendation, $r_S$ = .17, $p$ < .001.

In order to test whether or not individual differences in essentialist beliefs about gender moderate the relationship between age of identification and likelihood of recommendation for surgery, we will use a PROCESS macro in R to test whether likelihood of surgery recommendation differs depending on age of first identification, entering age of first identification as the predictor, and essentialist beliefs about gender and gender identity as moderators, and likelihood of surgery recommendation as the outcome.

Although both gender essentialism ($b$ = 8.97, $SE$ = 2.13, $p$ < .001).and target gender identity ($b$ = -16.78, $SE$ = 6.92, $p$ = .02) were significantly associated with likelihood of recommendation for surgery, we found no evidence that either of these characteristics moderated the relationship between age of identification and likelihood of surgery recommendation (age of identification x gender essentialism: $b$ = 1.25, $SE$ = 1.37, $p$ = .36; age of identification x target gender identity: $b$ = 5.43, $SE$ = 4.37, $p$ = .21)

***Full Output of All Mixed Effects Models Reported in Main Manuscript***

Mixed effects model: gender identity x gender expression (preregistered)

This regression model included three fixed effect predictors: gender identity (transgender woman [contrast-coded as 0.5] vs. non-binary person [contrast-coded as -0.5]), age of identification (early identification [contrast-coded as 0.5] vs. adolescent identification [contrast-coded as -0.5]), and the interaction between these two factors.

| | surgery_recommendation | | | |
|---|---|---|---|---|
| *Predictors* | *Estimates* | *std. Error* | *CI* | *p* |
| (Intercept) | 58.25 | 1.28 | 55.75 – 60.76 | **<0.001** |
| target gender modality | 12.17 | 2.55 | 7.16 – 17.19 | **<0.001** |
| age of first identification | -1.10 | 2.55 | -6.12 – 3.91 | 0.666 |
| target gender modality × age of first identification | 2.66 | 5.11 | -7.37 – 12.69 | 0.602 |
| Observations | 630 | | | |
| R2 / R2 adjusted | 0.036 / 0.031 | | | |

## Additional Information about Study 2[30]

***Detailed Method***

The goal of this study is to learn more about how people decide who should be prioritized gender-affirming healthcare. To do this, we asked two sets of questions. First, in order to understand what characteristics participants think are important to consider when deciding who to prioritize for gender affirming healthcare, participants were presented with the question *"To what extent do you think each of the following characteristics should be considered when determining whether a gender diverse adult (18+) should receive gender affirming surgery?"* Then, they were asked to rate several characteristics (included in the Table S1 below) on a scale from 0 (Should definitely not be considered) to 100 (Should definitely be considered). Then, to better understand what kinds of surgery candidates participants thought wanted gender affirming surgery, participants were asked to *"Use the sliders below, to rate how likely you think a gender diverse person with each of the following characteristics would want gender affirming healthcare on a scale from 0 (very unlikely to want gender affirming surgery) to 100 (very likely to want gender affirming surgery)?"* (for full list of characteristics, see Table S2 below).

***Full Analyses***

First, we ran a series of exploratory one-sample t-tests (comparing participants' ratings to a neutral value of 50) to see which characteristics (if any) participants rated as particularly important or unimportant when determining whether or not to recommend a person for gender affirming surgery (See Table S1 below and Figure 3 in manuscript).

**Table S1**

| Characteristic | M | 99.5% CI | *df* | *t* | *p* |
|---|---|---|---|---|---|
| A person's current age | 77.34 | 68.13 - 86.54 | 100 | 8.70 | <.001*** |
| A person's current gender identity | 49.75 | 39.07 - 60.44 | 100 | -0.07 | .95 |
| A person's age when they first realized they were gender diverse | 57.22 | 47.14 - 67.30 | 99 | 2.10 | .04* |
| Amount of time a person has been living as current gender identity | 63.27 | 54.03 - 72.50 | 100 | 4.21 | <.001*** |
| The extent to which a person presents like their current gender | 41.53 | 32.05 - 51.01 | 99 | -2.62 | .01* |
| How certain a person is that they want surgery | 82.12 | 73.50 - 90.73 | 100 | 10.92 | <.001*** |

[30] This study was not pre-registstered

| | | | | | |
|---|---|---|---|---|---|
| How long a person has wanted surgery | 69.56 | 60.73 - 78.40 | 100 | 6.49 | <.001*** |
| A person's mental health | 85.69 | 78.68 - 92.71 | 100 | 14.91 | <.001*** |
| The amount of social support a person has | 58.74 | 48.87 - 68.61 | 100 | 2.60 | .01* |
| A person's financial means | 55.97 | 45.72 - 66.23 | 99 | 1.71 | .09 |
| How far a person lives from the surgical center | 24.67 | 16.61 - 32.73 | 99 | -9.21 | <.001*** |
| A person's political orientation | 9.57 | 3.39 - 15.75 | 99 | -19.18 | <.001*** |

**Table S1.** *Exploratory one-sample t-tests looking at importance of various characteristics when considering a person for gender-affirming surgery (in comparison to a neutral value of 50).*

Additionally, we ran a second series of exploratory one-sample t-tests (again comparing participants' ratings to a neutral value of 50) to see which characteristics participants believed were associated with wanting affirming or not wanting gender affirming surgery (see Table S2 and Figure 4 in manuscript)

**Table S2**

| Characteristic | M | 99.7% CI | *df* | *t* | *p* |
|---|---|---|---|---|---|
| A person who identifies as transgender | 81.28 | 74.98 - 87.58 | 100 | 15.21 | <.001*** |
| A person who identifies as nonbinary | 43.58 | 35.15 - 52.01 | 100 | -2.33 | .02* |
| A person who identifies as transgender & nonbinary | 66.04 | 58.41 - 73.67 | 100 | 6.44 | <.001*** |
| A person who first recognized their gender diverse identity in early childhood | 74.00 | 67.59 - 88.41 | 100 | 11.47 | <.001*** |

| | | | | | |
|---|---|---|---|---|---|
| A person who first recognized their gender diverse identity in adolescence | 71.49 | 65.52 - 77.45 | 100 | 11.04 | <.001*** |
| A person who first recognized their gender diverse identity in young adulthood | 70.69 | 64.80 - 76.59 | 100 | 10.76 | <.001*** |
| A person who first recognized their gender diverse identity when they were middle aged | 59.22 | 51.93 - 66.51 | 100 | 3.88 | <.001*** |
| A person who first recognized their gender diverse identity in older adulthood | 53.20 | 44.99 - 61.41 | 100 | 1.19 | .24 |
| A person whose physical health is good | 65.43 | 57.16 - 73.69 | 100 | 5.72 | <.001*** |
| A person whose physical health is poor | 42.06 | 34.34 - 49.77 | 100 | -3.15 | .002** |
| A person whose mental health is good | 59.78 | 49.83 - 69.73 | 100 | 3.01 | .003** |
| A person whose mental health is poor | 59.92 | 50.66 - 69.18 | 100 | 3.28 | .001** |
| A person who is financially well off | 68.11 | 60.31 - 75.91 | 100 | 7.11 | <.001*** |
| A person who struggles to pay their bills | 45.29 | 36.58 - 54.00 | 100 | -1.66 | .10 |
| A person who has been waiting for surgery a long time | 87.71 | 82.64 - 92.79 | 100 | 22.76 | <.001*** |
| A person who has recently decided they want surgery | 59.94 | 52.33 - 67.55 | 100 | 4.00 | <.001*** |

| | | | | | |
|---|---|---|---|---|---|
| A person whose family & friends are supportive of their gender | 58.74 | 48.42 - 69.07 | 100 | 2.60 | .01* |
| A person whose family & friends are unsupportive of their gender | 50.87 | 42.68 - 59.06 | 100 | 0.33 | .75 |

Table S2. *Exploratory one-sample t-tests looking at the extent to which individuals with various characteristics are associated with wanting gender-affirming surgery (in comparison to a neutral value of 50).*

## Additional Information about Study 3

***Example Surgery Readiness Forms***

# Readiness for Gender Affirming Surgery Form

**Name:** Kai **Pronouns** she/her **Age (in years):** 20

**Gender Identity:** transgender woman **Birth Assigned Sex:** ☐ Female ☒ Male

**Current employment status (check all that apply):**

☐Full time work ☐Part time work ☒Student ☐Unemployed, looking for work
☐Unemployed, not looking for work ☐Currently on disability

**What kind of gender affirming surgery are you seeking? (check all that apply):**

☐Breast Augmentation ☐Facial Feminization Surgery ☐Electrolysis ☐Orchiectomy/testes removal
☐Tracheal shave/Adam's Apple reduction ☒Vaginoplasty/vagina construction

**Do you have someone you drive to and from your surgery appointment? Yes ☒ No☐**

**Do you have someone who will stay with you 24 hours after your surgery? Yes ☒ No☐**

**Please list all known allergies** strawberries

**Please list all sources of financial support you will rely on for surgery and post-operative care**

savings, part-time job, parents

**On a scale of 0 (not at all) 10 (very much), how much do you want gender-affirming surgery?**

0 1 2 3 4 5 6 7 8 (9) 10

**On a scale of 0 (not at all) 10 (very much), how important is it for you to get gender-affirming surgery?**

0 1 2 3 4 5 6 7 8 9 (10)

**For each of the questions below, please indicate the difficulty of each activity by using the following key and circling the most appropriate number:**

0 = No Difficulty
1 = Mild Difficulty
2 = Moderate Difficulty
3 = Severe Difficulty
4 = Extreme Difficulty or Cannot Do

**Following medical instructions?**

(0) 1 2 3 4

**Concentrating on daily activities?**

0 (1) 2 3 4

**Communicating with medical providers?**

0 (1) 2 3 4

**Maintaining your personal hygiene?**

(0) 1 2 3 4



Figure S1a. Example surgery readiness form of a transgender woman with 'high' desire

# Readiness for Gender Affirming Surgery Form

**Name:** Kai **Pronouns** she/her **Age (in years):** 20

**Gender Identity:** transgender woman **Birth Assigned Sex:** ☐ Female ☒ Male

**Current employment status (check all that apply):**

☐Full time work ☐Part time work ☒Student ☐Unemployed, looking for work
☐Unemployed, not looking for work ☐Currently on disability

**What kind of gender affirming surgery are you seeking? (check all that apply):**

☐Breast Augmentation ☐Facial Feminization Surgery ☐Electrolysis ☐Orchiectomy/testes removal
☐Tracheal shave/Adam's Apple reduction ☒Vaginoplasty/vagina construction

**Do you have someone you drive to and from your surgery appointment? Yes** ☒ **No** ☐

**Do you have someone who will stay with you 24 hours after your surgery? Yes** ☒ **No** ☐

**Please list all known allergies** strawberries

**Please list all sources of financial support you will rely on for surgery and post-operative care**

savings, part-time job, parents

**On a scale of 0 (not at all) 10 (very much), how much do you want gender-affirming surgery?**

0 1 2 3 4 5 (6) 7 8 9 10

**On a scale of 0 (not at all) 10 (very much), how important is it for you to get gender-affirming surgery?**

0 1 2 3 4 5 6 (7) 8 9 10

**For each of the questions below, please indicate the difficulty of each activity by using the following key and circling the most appropriate number:**

0 = No Difficulty
1 = Mild Difficulty
2 = Moderate Difficulty
3 = Severe Difficulty
4 = Extreme Difficulty or Cannot Do

**Following medical instructions?**

(0) 1 2 3 4

**Concentrating on daily activities?**

0 (1) 2 3 4

**Communicating with medical providers?**

0 (1) 2 3 4

**Maintaining your personal hygiene?**

(0) 1 2 3 4



Figure S1b. Example surgery readiness form of a transgender woman with 'regular' desire

# Readiness for Gender Affirming Surgery Form

**Name:** Kai **Pronouns** they/them **Age (in years):** 20

**Gender Identity:** nonbinary **Birth Assigned Sex:** ☐ Female ☒ Male

**Current employment status (check all that apply):**

☐Full time work ☐Part time work ☒Student ☐Unemployed, looking for work
☐Unemployed, not looking for work ☐Currently on disability

**What kind of gender affirming surgery are you seeking? (check all that apply):**

☐Breast Augmentation ☐Facial Feminization Surgery ☐Electrolysis ☐Orchiectomy/testes removal
☐Tracheal shave/Adam's Apple reduction ☒Vaginoplasty/vagina construction

**Do you have someone you drive to and from your surgery appointment? Yes ☒ No☐**

**Do you have someone who will stay with you 24 hours after your surgery? Yes ☒ No☐**

**Please list all known allergies** strawberries

**Please list all sources of financial support you will rely on for surgery and post-operative care**

savings, part-time job, parents

**On a scale of 0 (not at all) 10 (very much), how much do you want gender-affirming surgery?**

0 1 2 3 4 5 6 7 8 (9) 10

**On a scale of 0 (not at all) 10 (very much), how important is it for you to get gender-affirming surgery?**

0 1 2 3 4 5 6 7 8 9 (10)

**For each of the questions below, please indicate the difficulty of each activity by using the following key and circling the most appropriate number:**

0 = No Difficulty
1 = Mild Difficulty
2 = Moderate Difficulty
3 = Severe Difficulty
4 = Extreme Difficulty or Cannot Do

**Following medical instructions?**

(0) 1 2 3 4

**Concentrating on daily activities?**

0 (1) 2 3 4

**Communicating with medical providers?**

0 (1) 2 3 4

**Maintaining your personal hygiene?**

(0) 1 2 3 4



Figure S1c. Example surgery readiness form of a nonbinary person with 'high' desire

# Readiness for Gender Affirming Surgery Form

**Name:** Kai **Pronouns** they/them **Age (in years):** 20

**Gender Identity:** nonbinary **Birth Assigned Sex:** ☐ Female ☒ Male

**Current employment status (check all that apply):**

☐Full time work ☐Part time work ☒Student ☐Unemployed, looking for work
☐Unemployed, not looking for work ☐Currently on disability

**What kind of gender affirming surgery are you seeking? (check all that apply):**

☐Breast Augmentation ☐Facial Feminization Surgery ☐Electrolysis ☐Orchiectomy/testes removal
☐Tracheal shave/Adam's Apple reduction ☒Vaginoplasty/vagina construction

**Do you have someone you drive to and from your surgery appointment? Yes** ☒ **No** ☐

**Do you have someone who will stay with you 24 hours after your surgery? Yes** ☒ **No** ☐

**Please list all known allergies** strawberries

**Please list all sources of financial support you will rely on for surgery and post-operative care**

savings, part-time job, parents

**On a scale of 0 (not at all) 10 (very much), how much do you want gender-affirming surgery?**

0 1 2 3 4 5 (6) 7 8 9 10

**On a scale of 0 (not at all) 10 (very much), how important is it for you to get gender-affirming surgery?**

0 1 2 3 4 5 6 (7) 8 9 10

**For each of the questions below, please indicate the difficulty of each activity by using the following key and circling the most appropriate number:**

0 = No Difficulty
1 = Mild Difficulty
2 = Moderate Difficulty
3 = Severe Difficulty
4 = Extreme Difficulty or Cannot Do

**Following medical instructions?**

(0) 1 2 3 4

**Concentrating on daily activities?**

0 (1) 2 3 4

**Communicating with medical providers?**

0 (1) 2 3 4

**Maintaining your personal hygiene?**

(0) 1 2 3 4



Figure S1d. Example surgery readiness form of a nonbinary person with 'high' desire

***Preregistered Analyses***

First, we will run a mixed effects model to predict likelihood of surgery recommendation (listed as DV 1 above) from target gender, desire for surgery, and their interaction. To do this, we will use a mixed effects model (listed below) which will include fixed effects of target gender, desire for surgery, and target gender*desire for surgery. The model will also include a random intercept for subject id.

surgery_recommendation ~ target_gender*surgery_desire + (1| subject_id)

The results of this mixed effects model revealed that participants in Study 3 are more likely to recommend surgery candidates with a high desire for surgery than those who want surgery a regular amount ($b$ = 28.39, $SE$ = 1.29, $p$ < .001). Additionally, participants were more likely to recommend surgery seekers who were transgender women than nonbinary people ($b$ = 2.71, $SE$ = 1.29, $p$ < .036). Finally, there is no interaction between surgery seeker's surgery desire and gender; $b$ = 1.82, $SE$ = 2.59, $p$ = .481).

We will also do 2 post-hoc t-tests[1] comparing whether or not likelihood of recommendation differs for either trans people or for nonbinary people as a function of surgery desire.

For both trans and nonbinary surgery seekers, participants were more likely to recommend individuals with a high desire for surgery than individuals with a regular desire for surgery ($t$(386.04) = 9.42, $p$ < .001 for trans surgery seekers and $t$(383) = -8.83, $p$ < .001 for nonbinary surgery seekers).

Second, we will recode participants' responses DV #2 to determine which "type" of individual participants chose to recommend in a forced-choice paradigm where they could only recommend one out of the four individuals they read about for surgery. Given our 2x2 design, the 4 possible options are (1) trans woman/high desire; (2) trans woman/medium desire; (3) nonbinary person/high desire; (4) nonbinary person/medium desire. Then, we run a chi-square goodness of fit test to test whether or not the observed distribution of participants' selection rates differ from chance (25% likelihood) selecting each "type" person.

A chi-square goodness of fit test suggests that when forced to select only one surgery seeker to recommend for surgery, participants' are selecting people at rates that are significantly different from what would be expected by random chance ($X^2$ (3, $N$ = 199) = 162.65, $p$ < .001.

We will compare the coefficients between surgery desire and gender identity to determine if a person's gender identity or a person's stated desire for surgery has a greater influence on likelihood of recommending a person for gender affirming surgery. However, we do not have a hypothesis on which will influence a person more.

In our mixed model above, we find that the coefficient for surgery desire ($b$ = 28.39) is over 10 times greater than the coefficient for gender identity ($b$ = 2.71), suggesting that surgery desire has a greater influence on likelihood of recommendation for gender affirming surgery.

***Full Output of All Mixed Effects Models Reported in Main Manuscript***

Mixed effects model: gender identity x gender expression (preregistered)

We used a mixed effects regression that included a random intercept for participant, a fixed effect for gender identity (transgender woman [contrast-coded as 0.5] vs. non-binary person [contrast-coded as -0.5]), a fixed effect for target surgery desire (high desire [contrast-coded as 0.5] vs. regular desire [contrast-coded as -0.5]), and a fixed effect for the interaction of these two factors.

| | surgery_recommendation | | | |
|---|---|---|---|---|
| *Predictors* | *Estimates* | *std. Error* | *CI* | *p* |
| (Intercept) | 47.87 | 1.90 | 44.15 – 51.59 | **<0.001** |
| target gender | 2.71 | 1.29 | 0.17 – 5.25 | **0.036** |
| desire level | 28.39 | 1.29 | 25.86 – 30.93 | **<0.001** |
| target gender × desire level | 1.82 | 2.59 | -3.25 – 6.90 | 0.481 |
| **Random Effects** | | | | |
| σ2 | 332.70 | | | |
| τ00 PROLIFIC_PID | 631.45 | | | |
| ICC | 0.65 | | | |
| N PROLIFIC_PID | 199 | | | |
| Observations | 796 | | | |
| Marginal R2 / Conditional R2 | 0.175 / 0.715 | | | |

## Chapter 5: General Discussion

As a field, psychology has long operated under the default assumption that we think about and experience gender as a binary system composed of (cisgender) men and women. This dissertation challenges this default and demonstrates that our perceptions, conceptualizations, and attitudes about gender are much more nuanced than a 2-category system. By focusing on gender beyond the binary, this dissertation answers basic questions about how we perceive, think about, and understand androgynous and nonbinary individuals. The first section clarifies perceptions of androgynous faces, and also challenges the universality of a 2-category system. The next section explores how people think about nonbinary people as distinct from other gender groups, and begins to investigate the facilitators and limits of cisgender people's support for nonbinary individuals in everyday life. The final section explores how we make decisions about policies that involve nonbinary individuals. Together, this dissertation lays the groundwork for exploring psychology beyond the binary.

### Challenging the inevitability of binary face categorization

*Summary.* The first chapter of this dissertation builds on previous research in social categorization to explore how individuals perceive and categorize faces with a more intermediate (or androgynous) appearance. Using a novel continuum measure that allowed participants to place faces along a spectrum between male and female, my collaborators and I observed that cisgender participants tended to sort more clearly gendered faces into two distinct groups at the opposite ends of the spectrum participants but exhibited more variability and errors when categorizing faces with more androgynous appearances. However, when we expanded our participant pool to include transgender and nonbinary individuals, we found that these participants defaulted to a trimodal sorting pattern, placing the most androgynous faces toward

the middle of the spectrum. Additionally, we demonstrated that short-term exposure to a third gender category in a perceptual task can influence these categorization patterns. Overall, our findings reinforce prior work suggesting that perceptions of face gender are shaped by familiar gender labels, while also suggesting that there is flexibility within this system, including the ability to perceive more nuanced than two gender categories.

*Contributions to psychology.* While previous work found that participants can accurately categorize androgynous faces as being "in-between" masculine/male and feminine/female on average (Freeman et al., 2010), it remained unclear what contributed to the average. One possible explanation is that participants conceptualized androgynous faces as truly in-between masculine and feminine faces and thus placed them in the middle. By contrast, another possibility is that some of the people interpreted the androgynous face as male and others interpreted it as female, resulting in an average rating of "in-between". Data presented in section 1 of this dissertation does not support either of these possibilities, and instead finds that cisgender people's default judgements about androgynous faces are uniformly distributed across the entire spectrum from male to female, resulting in no clear categorization pattern for androgynous faces. By clarifying the distribution of cisgender participants' categorizations, we demonstrate that there is perhaps greater error when making judgments about androgynous faces than previously thought, but that these inaccuracies are not biased in a particular direction.

Our results also challenge assumptions about the inevitability and rigidity of binary gender categorization in at least two ways. First, collecting data from a sample of gender minority (trans and/or nonbinary) individuals who spontaneously categorized the faces into 3 groups (instead of two) demonstrates that not everyone shares a default binary structure when it comes to the categorization of gender/sex. Additionally, these findings also build upon and

extend 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 (Peery & Bodenhausen, 2008; Chen & Hamilton, 2012; Young et al., 2021; Chen et al., 2023; Levy et al., 2023). This dissertation extends these findings to the domain of gender/sex and demonstrates that when cisgender participants are given an intermediate (or 'third’) option their categorization of androgynous faces becomes more accurate. Although the exact way in which the third option is presented (with or without labels, etc.) influences its effectiveness, the mere fact that adding an additional option changes (cisgender) people’s categorizations challenges the rigidity of binary gender/sex categorization.

*Limitations and future directions.* While these studies provide novel insights about how individuals perceive and categorize androgynous faces, there are several limitations to this work which suggest future directions. First, our stimuli consisted of men’s and women’s faces that were systematically morphed at 5% and 10% increments along a continuum between male and female. In reality, androgynous faces are not combinations of men’s and women’s faces, raising the question of whether our findings would generalize to how people categorize faces that are more ecologically valid. Future research should explore how people categorize non-morphed faces that vary naturally along the dimensions of gender/sex to determine whether the same patterns hold outside an experimental context. In addition to the stimuli themselves, another limitation of our study was the structure of the continuum measure itself, which may have imposed a linear structure on participants’ thinking that does not fully capture the complexity of gender/sex perception. While a linear continuum between male and female is an improvement upon previous categorical measures, gender/sex does not exist in a strictly linear fashion. Thus, future research should consider whether alternative task structures—such as allowing

participants to sort faces along a two-dimensional plane—might produce different categorization patterns. Despite measuring several possible covariates, this work was unable to capture the underlying cognitive and social mechanisms that may underlie participants' performance on the categorization task. Thus, future work should look more carefully at other potential mechanisms that might play a role in how both cisgender and transgender individuals engage with gender categorization tasks. Finally, this work investigates one way that categorization tasks can be set up in such a way that induces cisgender people to add a third gender category. However, we do not explore the extent to which these effects persist across time and whether or not the change in categorization translates to other ideas about gender outside the experimental setting. Thus, future research should continue to explore how default categorization patterns for gender emerge and whether and how they can be shifted over time and in different contexts.

**Exploring conceptualizations of nonbinary people**

*Summary*. In the second section, this dissertation also takes steps toward understanding how the majority of people (who are cisgender) conceptualize and think about nonbinary people as a group. Building on prior research about 'intermediate' groups - those seen as existing between more prominent groups within the same identity dimension (Burke et al., 2023) - we first confirm that nonbinary people are often categorized as being ‘in-between’ traditional gender groups. Additionally, our results find that, similar to other intermediate groups, nonbinary individuals are viewed as less socially legitimate, less stable in their identities, and less likable than other gender groups. Lastly, we demonstrate that these perceptions have practical implications, as people’s views of nonbinary individuals predict their support for policies that are inclusive of nonbinary identities. Overall, our findings suggest that the perceived lack of

legitimacy and stability surrounding nonbinary identities may contribute to reduced societal support for affirming policies.

*Contributions to psychology*. This provides a unique contribution to past work which has assumed that people think about nonbinary people similarly to how they think about transgender people - or simply excluded consideration of nonbinary people at all. While ideas about transgender and nonbinary people may be somewhat overlapping (Perez-Arche & Miller, 2021), this work demonstrates that they are also distinct. This dissertation presents evidence to suggest that (cisgender) people think about nonbinary people as being less socially real than transgender people, as well as less stable and less likable when directly compared to transgender women and men. By focusing specifically on perceptions of nonbinary people, this dissertation begins to explore the ways we think about nonbinary people in relation to and as distinct from those of other gender groups.

*Limitations and future directions.* This line of work begins to explore how cisgender individuals think about nonbinary people in relation to more well-known gender groups, and leaves several avenues for future work. For instance, while we establish that participants think about nonbinary individuals as "in-between" men and women, it remains unclear whether participants conceptualize nonbinary individuals as forming a singular cohesive group, as a more heterogeneous set of individuals spread out across the a spectrum between male and female, or even as as a series of subgroups occupying various positions along the gender spectrum. Additionally, while we were specifically focused on perceptions of nonbinary people, our research does not rule out the possibility that other groups—particularly transgender people—might also be perceived as “in-between” gender categories by some participants. This raises the question of whether the perception of being “in-between” is truly unique to nonbinary

individuals or if it might apply to other gender minorities as well. Thus, future research could more systematically explore what it means to think of nonbinary people as "in-between" other gender groups, and the extent to which ideas about intermediateness might also apply to other gender minorities such as transgender people. Another limitation of this work is that the correlational nature of our data precludes our capacity to make causal and/or directional inferences about the relationship between contact with nonbinary individuals and support for nonbinary-affirming policies. While our data suggest that greater contact with nonbinary individuals is associated with greater support for their rights, it remains unclear whether it's the case that greater contact leads to increased support, or whether individuals who already hold more supportive views are more likely to have contact with nonbinary people. Thus, future research would benefit from experimental designs that include a condition with direct contact or exposure to a nonbinary person in order to more rigorously test the causal role of interpersonal contact in shaping policy attitudes toward nonbinary individuals. Finally, despite the goal of better understanding how we think about nonbinary people, our research not explore how nonbinary people think of themselves. Thus, future research that includes the perspectives of nonbinary participants would offer valuable insights on how nonbinary people understand themselves in relation to other gender groups, while also providing a more nuanced understanding of how these individuals navigate societal perceptions and how their self-perception interacts with the biases they face.

**Investigating facilitators and limitations of nonbinary affirmation and support**

*Summary*. In section 3, we investigate the factors associated with support for nonbinary affirming policies in relation to access to gender-affirming healthcare. Specifically, we focus on how cisgender people prioritize candidates for gender-affirming surgery when demand exceeds

supply. Our findings show that cisgender people tend to prioritize those whose gender identities and expressions fit more binary narratives—such as binary trans individuals, people who came out at younger ages, and those with more distinctly binary presentations. To understand this pattern, we find that binary trans people and those who transitioned earlier in life are seen as more likely to desire surgery compared to nonbinary individuals and those who transitioned later. Even when the level of desire for surgery is made equivalent, nonbinary individuals are still deprioritized in surgical recommendations. Together these findings suggest that people's assumptions about who seeks gender-affirming care influence decisions about who should receive it, which often leads to the marginalization of nonbinary people.

*Contributions to psychology*. While past literature has focused on support for transgender affirming policies (Barbir et al., 2017; Reimer et al., 2017; Miller et al., 2020; Morgenroth et al., 2024), considerably less has focused on support for nonbinary people, which is distinct in several ways. While those advocating for (binary) trans people's rights may make the case that trans rights are a necessary for trans people to better fit within the gender binary system and have the "same rights" as (cis) men and women, those advocating for nonbinary people's rights can't appeal to consistency, and instead must justify supporting rights for individuals who exist outside the binary system. Perhaps unsurprisingly, participants in our studies showed meaningful differences in their willingness to support nonbinary vs. trans people, particularly when it comes to their willingness to recommend them for gender affirming surgery. Here, we found repeated evidence that people were more likely to recommend transgender over nonbinary people for gender affirming surgery, even when desire for surgery was made equivalent between transgender and nonbinary surgery seekers. Thus, we begin to uncover factors that facilitate and limit cisgender people's support for nonbinary affirming policies.

*Limitations and future directions*. These findings begin to shed light on how cisgender individuals make decisions about gender affirming policies in the context gender-affirming healthcare, and leaves open several avenues for future research. First, this set of studies focused exclusively on healthcare relevant to transfeminine and nonbinary individuals assigned male at birth. Thus, it is unclear whether or not cisgender individuals rely on the same set of criterion when making decisions about gender-affirming care for trans men and nonbinary individuals assigned female at birth. Existing work suggests that while there is some overlap, cisgender people have different concerns about trans men and women when it comes to trans-inclusive policies. Thus, future studies in this line of work should explore whether cisgender people rely on similar types of criteria when prioritizing people for gender-affirming surgeries and procedures relevant to individuals assigned female at birth. Another important limitation is that we did not sample healthcare providers, who are the most critical decision-makers in determining access to gender-affirming care. As a result, it remains unclear whether healthcare providers share the proclivities of other cisgender people when it comes to deciding who to prioritize for gender-affirming surgery, or if their professional training and experience would lead them to make different decisions. Future work looking directly at healthcare providers' decisions could provide more practical and policy-relevant insights into how decisions around gender-affirming care are made.

In this set of studies, we find in separate studies that (1) people believe transgender women want surgery more than nonbinary individuals and (2) that people are more likely to recommend trans women for surgery than nonbinary people for surgery. However, future work should replicate both of these findings within a single experimental framework to directly test whether perceptions of greater desire for surgery are directly related to the prioritization of trans

women over nonbinary candidates for gender affirming care. In the same vein, our results suggest that cisgender participants are more likely to recommend trans women than nonbinary individuals for surgery even when desire for surgery is equal. This finding suggests the existence of additional reasons beyond surgery candidates' perceived desire that may motivate participants' prioritization for gender affirming surgery. Thus, future research should explore the potential mediating role of other possible factors such as identity legitimacy, conformity to societal norms, or perceptions of deservingness in participants' tendency to prioritize trans women over nonbinary people for surgery. Understanding these motivations could provide key insights into how and why nonbinary people are consistently marginalized in the context of gender-affirming care and offer potential pathways for interventions to reduce these biases.

**General Limitations and Future Directions**

In addition to future directions specific to each line of research, there are some additional limitations that apply broadly to the work presented in this dissertation that suggest future directions for the study of gender diversity. First, the data in these studies were gathered from convenience samples of online participant pools and college student pools. Past work has asserted that relying on convenience samples tends to result in an overrepresentation of individuals who are Western, educated, rich and industrialized (see Henrich et al., 2010; Doebel & Frank, 2024). As a result, the findings from our studies may not generalize beyond a narrow slice of humanity and are not a reliable basis from which to generate general theories about gender diversity. Additionally, the results do not capture the ways in which other relevant sociocultural characteristics such as race (see Johnson, Freeman, & Pauker, 2012; Lei, Foster-Hanson, & Goh, 2023), and cultural background (see Best & Williams, 2001) may intersect with ideas about gender diversity. Thus, future work should intentionally seek

perspectives from a more representative sample of individuals in order to achieve a more holistic picture of existing ideas about gender diversity.

Additionally, this dissertation focused primarily on investigating *cisgender* people's perceptions of nonbinary people and faces with androgynous gender appearance. While this is a reasonable place to start, it falls short of being able to answer questions about gender diversity in (at least) two important ways. First, the studies in this dissertation ask specifically about transgender and/or nonbinary people (papers 2 and 3), and people with androgynous appearances (paper 1). However, there are many individuals who do not fall into any of these categories but may still self-identify as being outside of the gender binary. Thus, future work should expand the operationalizations of what it means to be "outside of the gender binary" to include other (combinations) of identities that do not coincide with a binary gender system such as genderfluid, agender, nonbinary demi-boy, etc.. In addition to a limited operationalization of gender diversity, this dissertation does not include the perspectives of gender diverse individuals (with the exception of Paper 1, Study 3). Past work (including the results of Paper 1, Study 3 of this dissertation) has suggested that people who identify as gender diverse may have different understandings of gender diversity than cisgender people (see also Nagoshi et al., 2012; Bradford et al., 2019; Factor & Rothblum, 2008; Barsigian et al., 2020). Thus, future research into how gender diverse people think about or understand their own (and others') gender identities is crucial to obtain a full understanding of gender outside of the binary.

In addition to limitations on "who" and "what" was studied, there were notable limitations in the methodology of the studies presented in this dissertation. Specifically, measures of gender perception were all cross-sectional, quantitative, and measured a single aspect of how people think about gender beyond the binary. In reality, people's experience of gender is

continuous, dynamic, and multi-faceted. Future research should focus on broadening our toolkit to capture more than the attitudes, perceptions, and impressions that can be collected and quantified using a single online survey or task in order to gain an understanding of how people experience gender in everyday life. For instance, the use of repeated sampling strategies like Ecological Momentary Assessment (Shiffman et al., 2008) or Electronically Activated Recorder (Mehl et al., 2001) will allow researchers to investigate how individuals of all genders experience gender in natural environments. Another approach to examining gender in context is to utilize large, existing archival datasets to explore how gender (diversity) has been discussed and represented over time (Charlesworth et al., 2021). Additionally, because gender is a multi-dimensional experience, future research could benefit from taking a mixed methods approach, using complementary qualitative and quantitative methods to gain a more nuanced picture of how people experience their own and others' gender.

**Conclusion**

Gender is one of the fundamental axes through which we think about ourselves and each other. This dissertation begins to demonstrate some of the ways that the field of psychology stands to gain from thinking about gender beyond the binary. By accounting for the existence of gender diverse individuals, this dissertation expands our knowledge of basic psychology and improves the accuracy of our mental representation of gender at the individual and group levels. Additionally, it uncovers how cisgender people conceptualize and evaluate nonbinary people as distinct from other gender groups, and explores the practical consequences of these attitudes—particularly in relation to support for policies affecting gender-diverse populations. In doing so, this dissertation advances the conversation on societal attitudes toward nonbinary people.

In addition to the specific contributions of the studies presented in this dissertation, a clearer understanding of gender diversity will simultaneously further broad goals in psychology beyond the scope of this dissertation such as improving our measures and enhancing our knowledge of other topics such as prejudice and discrimination, intergroup relationships. Additionally, a diverse perspective on gender leaves us open to the ways our understandings of gender/sex will inevitably grow and expand in the coming years. Even if the specific findings of this dissertation become obsolete as our understanding continues to evolve, embracing a framework that considers gender beyond the binary leaves us poised to explore and understand future conceptualizations and experiences of gender as they arise.

## References


Barbir, L. A., Vandevender, A. W., & Cohn, T. J. (2017). Friendship, attitudes, and behavioral intentions of cisgender heterosexuals toward transgender individuals. *Journal of Gay & Lesbian Mental Health*, *21*(2), 154–170. https://doi.org/10.1080/19359705.2016.1273157

Barsigian, L. L., Hammack, P. L., Morrow, Q. J., Wilson, B. D. M., & Russell, S. T. (2020). Narratives of gender, sexuality, and community in three generations of genderqueer sexual minorities. *Psychology of Sexual Orientation and Gender Diversity*, *7*(3), 276–292. https://doi.org/10.1037/sgd0000384

Best, D. L., & Puzio, A. R. (2019). Gender and Culture. In D. Matsumoto & H. C. Hwang (Eds.), *The Handbook of Culture and Psychology* (2nd ed., pp. 235–291). Oxford University PressNew York. https://doi.org/10.1093/oso/9780190679743.003.0009

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. *The International Journal of Transgenderism*, *20*(2–3), 155–168. https://doi.org/10.1080/15532739.2018.1474516

Burke, S. E., Perry, S. P., Dovidio, J. F., & LaFrance, M. (2023). Distinctive negative reactions to intermediate social groups. *Journal of Applied Social Psychology*, *53*(7), 600–630. https://doi.org/10.1111/jasp.12942

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words. *Psychological Science*, *32*(2), 218–240. https://doi.org/10.1177/0956797620963619

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*. https://doi.org/10.1177/01461672231190264

Doebel, S., & Frank, M. C. (2024). Broadening Convenience Samples to Advance Theoretical Progress and Avoid Bias in Developmental Science. *Journal of Cognition and Development*, *25*(2), 261–272. https://doi.org/10.1080/15248372.2023.2270055

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

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

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology*, *102*(1), 116–131. https://doi.org/10.1037/a0025335

Lei, R. F., Foster-Hanson, E., & Goh, J. X. (2023). A sociohistorical model of intersectional social category prototypes. *Nature Reviews Psychology*, *2*(5), 297–308. https://doi.org/10.1038/s44159-023-00165-0

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

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, *33*(4), 517–523. https://doi.org/10.3758/BF03195410

Miller, P. R., Flores, A. R., Haider-Markel, D. P., Lewis, D. C., Tadlock, B., & Taylor, J. K. (2020). The Politics of Being "Cait": Caitlyn Jenner, Transphobia, and Parasocial Contact Effects on Transgender-Related Political Attitudes. *American Politics Research*, *48*(5), 622–634. https://doi.org/10.1177/1532673X20906460

Morgenroth, T., Axt, J. R., & Westgate, E. C. (2022). What Underlies the Opposition to Trans-Inclusive Policies? The Role of Concerns About Male Violence Versus Attitudes Toward Trans People. *Personality and Social Psychology Bulletin*, 014616722211372. https://doi.org/10.1177/01461672221137201

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

Peery, D., & Bodenhausen, G. V. (2008). Black + White = Black: Hypodescent in Reflexive Categorization of Racially Ambiguous Faces. *Psychological Science*, *19*(10), 973–977. https://doi.org/10.1111/j.1467-9280.2008.02185.x

Perez-Arche, H., & Miller, D. J. (2021). What Predicts Attitudes toward Transgender and Nonbinary People? An Exploration of Gender, Authoritarianism, Social Dominance, and Gender Ideology. *Sex Roles*, *85*(3–4), 172–189. https://doi.org/10.1007/s11199-020-01212-3

Reimer, N. K., Becker, J. C., Benz, A., Christ, O., Dhont, K., Klocke, U., Neji, S., Rychlowska, M., Schmid, K., & Hewstone, M. (2017). Intergroup Contact and Social Change: Implications of Negative and Positive Contact for Collective Action in Advantaged and Disadvantaged Groups. *Personality and Social Psychology Bulletin*, *43*(1), 121–136. https://doi.org/10.1177/0146167216676478

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological Momentary Assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

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

**Acknowledgements**

For an intensely private person brimming with gratitude to all those who have shepherded me to this point, anything written here feels simultaneously too little and too much.

First, I'd like to thank the academic mentors who have encouraged, inspired, guided, invested in, and advocated for me over the past decade.

Thank you to Cindy for being the first person to make me excited about psychology. Thank you for taking me on as a driven-but-clueless undergraduate when I showed up in your Social Psychology class, for giving me a research home in your lab, for taking me to my first conference, and for supporting me through my undergraduate thesis.

Thank you to Sam for taking me on as your lab manager, for taking my research interests seriously during a time when I doubted whether or not they had a place in the academy, and for doing everything in your power to set me up for success in graduate school.

Thank you to Betsy for playing a significant role in my decision to come to Princeton, and for five years of rigorous, interdisciplinary, and methodologically diverse lab meetings. I consistently come away from time spent with you and your research group brimming with big ideas.

Thank you to Diana for welcoming me into your lab as your secondary advisee early on in my time at Princeton. As a new graduate student finding my footing in a fully remote department during the early years of COVID-19, interacting with you and your lab was a highlight and something I looked forward to each week.

Thank you to Thekla for being a mentor personally and professionally, and for generally going above and beyond to help me out. Not only were you a reader on my dissertation, but you've also been a collaborator, letter writer, conference co-organizer, and (maybe most importantly) one of the first people to show me that I didn't have to "choose" between being a scholar and being nonbinary.

Kristina, I don't have words to describe how profound it has been to have your support. When I describe my experience working with you to others, I always say that it feels like you're helping me become the best version of myself. Thank you for having me at your house for six consecutive Thanksgiving dinners, for showing up at my house with a week's worth of food during the pandemic after I had major surgery, and for being on my team for the past five and a half years and beyond.

In addition to academic mentors, I want to express my gratitude to Beth, Betsy, Casey, Cheryl, David, Jeanny, Jenn, Jill, Ken, Rosemarie, Todd, and Yael for keeping the department running

and for all your efforts behind the scenes to advocate for graduate students. Thank you also to all the phenomenal individuals I've had the pleasure of crossing paths with as research staff during graduate school. Chase, Courts, Daniel, Grey, Jordan, Mia, MJ, Nat, Rachel, Shira, and Tori—thank you for keeping the lab running and making the work that the rest of us do possible.

At the beginning of graduate school, I told myself not to make friends because I thought I might leave, but several PhDs and one wedding(!) later, I couldn't be more delighted to say that I failed. Ella, Jessica, Kim, Léïla, and Tabea—thank you for choosing to stay in my life far beyond when I left Washington, and for always welcoming me back so warmly. We may be in very different places than we'd imagined, and I can't wait to watch where we go next.

To the friends I've made since coming to Princeton, thank you for making this place feel like home. Abla, Alex, Benny, Kara, Lauren, and Natalie—I couldn't be more grateful to have you all around to share in the joyful, excruciating, and mundane moments of life. You all have given me a masterclass in how to give and receive love and care.

Thanks also for letting me befriend and care for all of your pets. To Athena, Juniper, Petunia, Phineas, Sunny, Tofu, and Willow—thank you for being the best emotional support animals I could ask for. I am grateful to each one of you for enduring my many kisses and all the times I've put my face in your fur without asking. To my own cat, Mouse, I love you more than you could ever know—thanks for being my motivation to get up each morning (to feed you) and work hard each day (to give you a better life).

To Cynthia, Elena, Emma, and Jo, who knew me long before I was stats and/or a graduate student, thank you for reminding me of where I've been and for sticking around long enough to see who I've become. Also, thanks for being patient when I'm bad at communicating and for keeping me grounded at times when life feels unmooring.

Kyzen, thank you for being what has felt like an unending well of support, especially during the past couple of months, for being my emergency contact, for learning and growing with me, and for seeing all the best and worst versions of myself and still choosing me. I am so proud to be your partner and can't wait to explore more of this world with you.

妈妈爸爸, thank you for setting my life in motion. I hope I've lived beyond your wildest dreams.