# Assessing implicit attitudes about androgyny☆

S. Atwood [a,*], Jordan R. Axt [b,c]

[a] Department of Psychology, Princeton University, Princeton, NJ 08540, USA
[b] Department of Psychology, McGill University, Montreal, QC H3A 1G1, Canada
[c] Project Implicit, Boston, MA 02109, USA

## ARTICLE INFO

## ABSTRACT

This research examines attitudes towards androgyny using a novel Implicit Association Test (IAT) that assesses implicit evaluations of gender conforming people (i.e., those who look stereotypically male or female) vs. androgynous people (i.e., those whose appearance includes a combination of masculine and feminine traits). Over 6 studies (*N* > 6000), we develop a gender expression IAT and present evidence for its internal validity and incremental predictive validity with relevant psychosocial attitudes, such as need for closure, political ideology, and support for nonbinary affirming policies. Although the IAT consistently revealed more positive associations towards gender conforming than androgynous people and was reliably correlated with parallel measures of explicit attitudes, it failed to predict several behavioral outcomes related to gender expression in contexts like judgment, perceptual fluency, and mouse-tracking. We discuss the implications of these results concerning the study of gender expression and implicit social cognition.

The term "androgyny" is characterized by a combination of masculine and feminine attributes, traits, and/or characteristics and may refer to assigned sex, gender identity, gender expression, or sexual identity (Nowak & Denes, 2016). Manifestations of human androgyny have existed in cultures across the world beginning as early as ancient Sumer, where some intersex men took on the role of priests, wore traditionally female clothing, and spoke in a traditionally female dialect (Leick, 2003). While androgyny can be an indicator of assigned sex or sexuality, the current work focuses primarily on the implications of androgynous gender expression (typically characterized by having a combination of phenotypically masculine and feminine characteristics). In particular, we explore the implications of androgynous gender expression for individuals who are gender diverse (i.e. those whose internal experience of gender identity falls outside of the binary framework of man/male or woman/female; Richards et al., 2016; Hegarty, Ansara, & Barker, 2018; Matsuno & Budge, 2017). At the same time, it is important to note that there is no "correct" way express gender diversity, and it is not necessary to look androgynous to have a gender diverse identity. However, due to gross conflation of gender expression and gender identity (Valdes, 1994), it is likely that preferences for androgyny relate to attitudes about gender diverse individuals in a similar way that preferences for individuals who look stereotypically feminine relate to attitudes towards cisgender women (Osmond & Martin, 1975; Swim & Cohen, 1997).

While evidence of gender diversity can be found as far back as ancient Greece and Rome (Surtees & Dyer, 2020), emerging evidence indicates that greater numbers of people in Western society are publicly embracing gender diverse identities (e.g. transgender, nonbinary, genderfluid, genderqueer, agender, and many others) for the first time (Clark et al., 2014; Perez-Brumer, Day, Russell, & Hatzenbuehler, 2017; Rider, McMorris, Gower, Coleman, & Eisenberg, 2018; Wilson, Choi, Herman, Becker, & Conron, 2017). Indeed, a recent poll by Pew Research Center found that 18% of US adults know somebody who uses gender neutral pronouns (e.g. they/them), and 52% reported being at least somewhat comfortable using gender neutral pronouns to refer to others (Geiger & Graf, 2019). The increase in people who identify outside of a binary gender system has also brought a shift in the recognition of gender diversity in the media, legal, and social spheres. For example, the past half decade alone has brought the addition of singular 'they' in Merriam Webster dictionaries (Schmidt, 2019), the first transgender superhero on TV (Romo, 2018), and legal protection of transgender identities at the federal level (Bostock v. Clayton County, 2020).

## 1. Perceptions of androgynous faces

Despite broader recognition of gender diversity in science and society, limited work has been done on the perceptions and attitudes towards individuals with androgynous appearances. However, the existing work suggests that gender perception is guided by expectations of a binary gender distribution, and that individuals who violate that expectation receive more negative evaluations. For example, early work in face categorization found that people perceive gender categorically even among targets whose phenotypic appearance is continuous (Campanella, Chrysochoos, & Bruyer, 2001). Similarly, a second research group found that when asked to give subjective gender judgments on faces that ranged in appearance from an extremely male to extremely female phenotype, participants' responses were nonlinear, despite the monotonic nature of the face stimuli. Additionally, a comparison of the absolute value between the objectively expected response based on face phenotype and participants' subjective responses revealed less accurate evaluations of androgynous faces vs gender typical faces, further suggesting a categorical perception of gender (Freeman, Rule, Adams, & Ambady, 2010).

More recently, Stern and Rule (2018) found that more physically androgynous (vs. sex-typical) transgender individuals were evaluated more negatively due to the increased effort required to identify the individual's sex. That is, perceivers appeared to have a need to categorize faces as clearly male or female, and stimuli that complicated such a categorization received more negative evaluations. Finally, a parallel line of work revealed that people evaluated and remembered androgynous faces as less gender congruent when those same faces were labeled as transgender versus cisgender (Wittlin, Dovidio, LaFrance, & Burke, 2018). Such biases appear to be at least partly driven by top-down expectations about transgender people, as one series of studies discovered that participants represented people as less gender-typical after being told that a person identifies as transgender compared to cisgender (Howansky, Albuja, & Cole, 2020).

## 2. Understanding variation in attitudes towards gender diverse individuals

While the previously reviewed findings suggest a preference for gender conformity over androgyny, there is reason to believe that the beliefs associated with these preferences vary considerably. For example, a nationally representative sample of heterosexual individuals found that attitudes towards transgender people were positively correlated with feelings towards lesbian, gay, and bisexual people as well as personal contact with sexual minorities, but were negatively correlated with measures of authoritarianism, anti-egalitarianism, and religiosity (Norton & Herek, 2013). More recent work has found that masculine self-identification and endorsement of biological gender essentialism (i. e. believing that gender is innate) were associated with lower levels of support for both women and transgender individuals (Harrison & Michelson, 2019; Wilton et al., 2019). Additionally, a cross-sectional study of cisgender heterosexual college students found that having at least one transgender friend was associated with decreased prejudice as well as greater support and acceptance of transgender people (Barbir, Vandevender, & Cohn, 2017). However, little research has been done to see what predicts attitudes towards lesser-known gender minorities (such as nonbinary individuals). In the present work, we investigate this question by examining how performance on self-report as well as behavioral measures is related to implicit and explicit attitudes towards individuals with an androgynous appearance.

## 3. The role of implicit attitudes

Prior research has revealed how expectations about a binary gender system influence our perceptions and behavior towards others. To date, however, attempts to understand variance in the judgment and perception of gender diverse individuals have relied largely on measures of self-report. These studies, while informative, overlook the potential contribution of implicit attitudes, that is, associations in mind that are less intentional and less influenced by conscious goals compared to explicit self-report measures (Moors & De Houwer, 2006).

Though there are many indirect measures that have been developed to assess implicit attitudes, the most popular remains the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). In an IAT, participants sort stimuli associated with an attribute (e.g., words that are either positive or negative) as well as stimuli associated with two categories (e.g., images of Black and White people) as quickly as possible as they appear one at a time. In critical blocks, participants categorize items from all four sets of stimuli (e.g., positive words, negative words, Black people, and White people) using two keys (e.g., categorizing White people and positive words with one key, Black people and negative words with the other key). In other critical blocks, the pairing between words and images is switched. The standardized difference in reaction time between the two forms of critical blocks is then a measure of the strength of implicit associations between those categories and attributes in memory.

The past 20 years of research on the IAT has shown that performance on the task consistently reveals intergroup biases in such implicit evaluations (Nosek et al., 2007). Moreover, the IAT has been shown to demonstrate incremental predictive validity relative to parallel measures of explicit (i.e., self-reported) attitudes, meaning the IAT can predict variance in outcomes even after controlling for the role of explicit evaluations. This incremental predictive validity has been demonstrated across intergroup behavior more generally (Buttrick et al., 2020; Kurdi et al., 2019) and within the context of gender specifically, as a recent study found that a novel IAT of implicit attitudes towards transgender versus cisgender people incrementally predicted outcomes like contact with transgender people and support for policies like the right for transgender people to serve in the military (Axt, Conway, Westgate, & Buttrick, 2021).

In light of past work showing the utility of implicit measures in understanding intergroup behavior, the present work sought to develop an IAT measuring implicit attitudes towards individuals who look stereotypically male or female (henceforth gender conforming people) versus individuals who have a combination of masculine and feminine traits (henceforth androgynous people), and uses the task to 1) investigate the magnitude of any biases in implicit evaluations towards androgynous people and 2) understand how such attitudes may predict other outcomes associated with gender diversity, such as support for specific policies or behavior towards targets varying in levels of gender conformity. As a result, these efforts will both expand the scope of work in implicit attitudes into a growing area of social psychological research, and allow researchers the ability to use insights from the field of implicit social cognition to better understand attitudes towards androgyny (Bar-Anan & Vianello, 2018; Nosek & Smyth, 2007).

Across six studies, we introduce a novel IAT measuring implicit evaluations of gender conforming people. First we established internal reliability and predictive validity across both text and image based versions (Study 1a) and using multiple stimuli sets (Study 1b). Next, in Study 2, we used structural equation modeling analyses to establish incremental predictive validity of the IAT over parallel explicit preferences. Then, we ran 3 high-powered studies (Studies 3a-3c) testing the IAT's ability to predict performance on behavioral tasks. Finally, we review our results within the context of existing literature, identify some limitations of our findings, and discuss implications for future work. For all studies, we report all measures, manipulations and exclusions.

## 4. Study 1a

Study 1a compared the validity of an image-based and text-based IAT measuring implicit evaluations of gender conforming vs. androgynous people. Specifically, the two IATs were compared on internal reliability,

magnitude of preferences for gender-conforming people, and strength of predictive validity concerning 1) explicit attitudes toward androgyny, 2) essentialist beliefs about gender and sex, and 3) tolerance of ambiguity.

### 4.1. Method

#### 4.1.1. Participants

A total of 830 volunteers completed the study through the Project Implicit research pool (https://implicit.harvard.edu).[1] Data from 21 participants were excluded based on the predetermined exclusion criteria of completing more than 10% of trials faster than 300 milliseconds on the IAT (a sign of careless responding; Nosek et al., 2007) for a final sample of 809 eligible participants (see Appendix A in the supplementary material for demographic breakdown of all studies). A total of 400 participants completed the word-based IAT and 409 completed the image-based IAT, which gave us over 80% power to detect a correlation of $r = 0.14$ within each condition and over 80% power to detect an effect size of Cohen's $q = 0.20$ when comparing correlations between conditions. See https://osf.io/fxc2m/ for Study 1a's pre-registration.

#### 4.1.2. Measures

##### 4.1.2.1. Implicit gender expression attitudes.
Each participant completed either an image version or a text-only version of the IAT. Both versions were designed to measure associations between androgynous and gender conforming individuals with positive and negative words. Positive words included "Happy", "Fantastic", "Joyous", "Cherish", "Excellent", "Delightful", "Terrific", & "Adore", and negative words included "Pain", "Disaster", "Awful", "Grief", "Disgust", "Horrific", "Nasty" & "Detest". Both IATs used category labels of "Androgynous" and "Male/Female" to represent gender conformity, and each IAT followed the standard seven-block design recommended in Nosek et al. (2007). We calculated IAT scores using the *D* algorithm (Greenwald, Nosek, & Banaji, 2003), which we computed by first finding the mean latency for each test block (3, 4, 6, and 7) along with two standard deviations (one for trials in Blocks 3 and 6, and another for trials in Blocks 4 and 7). Then, we computed two mean difference scores (one between blocks 3 and 6 and the other between blocks 4 and 7), divided each difference score by its associated standard deviation, and took the average of the two quotients. For all participants, higher values indicated more positive associations with gender conforming versus androgynous people.

Stimuli in the text-based IAT consisted of words related to gender conforming people ("Gender Conforming", "Male", and "Female") and androgynous people ("Androgynous", "Gender Non-Conformity", and "Gender Non-Conforming People"). Stimuli in the the image IAT consisted of eight headshots (four male/female, four androgynous) matched on race. All male/female images were taken from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015). Androgynous images were created by morphing one of the male/female headshots plus another headshot of an individual (also from the Chicago Face Database) of the same race but "opposite" gender (see Appendix B in the supplementary material for stimuli used in image IAT and https://osf.io/64egx/ for stimuli from all studies).[2]

Since the distinction between gender conformity vs androgyny may not be salient in the everyday lives of our participants, we included the

---

[1] All data were collected in 2020.

[2] We opted to use morphed faces over faces of gender diverse individuals as stimuli in order to be consistent with prior work on gender face perception (e.g. Campanella et al., 2001; Freeman et al., 2010). Additionally, since there are no collections of faces from a gender diverse population that have been standardized for use in research, using photos from the Chicago Face Database (Ma et al., 2015) allowed us to control for image characteristics (e.g. camera angle, luminosity, background, face size, etc) that may be more difficult to standardize across a collection of naturalistic photos.

following disambiguation as part of the instructions: "For the purposes of this study, 'Androgynous' refers to people who exhibit both masculine AND feminine attributes and are not easily labeled as masculine or feminine. Conversely, the term 'Male/Female' refers to people who exhibit either masculine OR feminine attributes and are easily labeled as masculine OR feminine." Additionally, due to the novel stimuli, the image version of the IAT had an additional 24-trial training block to solidify the association between each of the eight images and its relevant category (see Axt et al., 2021).

##### 4.1.2.2. Explicit attitudes towards gender conformity/androgyny.
Explicit gender identity attitudes were measured using five items. First, participants reported their relative preference between gender conforming people (male/female) and androgynous people on a scale from 1 = *I strongly prefer androgynous people to male/female people* to 7 = *I strongly prefer male/female people to androgynous people* (Axt, 2018). Next, participants separarely rated warmth towards gender conforming and androgynous people using two pairs of feelings thermometers (see supplementary materials available at https://osf.io/yvdkx/ for full text). To mirror the IAT score, we created separate difference scores for the warmth and positivity, such that higher values indicated greater preference towards gender conforming people. These two difference scores and the relative preference item were standardized and averaged to create a composite explicit attitude item (α = 0.80; see Axt, Bar-Anan, & Vianello, 2020; Buttrick et al., 2020 for similar strategy).

##### 4.1.2.3. Gender/sex diversity beliefs scale (GSDB).
Participants responded to a 24-item Gender/Sex Diversity Beliefs Scale (GSDB) which measured essentialist and nonessentialist beliefs about gender & sex across 5 factors of Affirmation (α = 0.92), Gender Normativity (α = 0.68), Uniformity (α = 0.5), Surgery (α = 0.89), and Upbringing (α = 0.63; Schudson & van Anders, 2021). All items were presented in a random order on a 7-point Likert response scale ranging from 1 = *Strongly disagree* to 7 = *Strongly agree*.

##### 4.1.2.4. Tolerance/Intolerance of Ambiguity Scale.
Participants reported their comfort with ambiguous situations using a 16-item scale measuring both their tolerance (α = 0.57) and intolerance (α = 0.58) of ambiguity (Budner, 1962). All items were presented in a random order on a 7-point Likert response scale ranging from 1 = *Strongly disagree* to 7 = *Strongly agree*.

#### 4.1.3. Procedure
Participants were randomly assigned to complete either the text or image based IAT and the self-report measures in a randomized order.

### 4.2. Results

We first compared the image and text IATs on internal reliability and *D* scores. We computed Cronbach's alpha (Cronbach & Meehl, 1955) for each IAT by matching the 60 critical trials in blocks 3–4 with the 60 critical trials in blocks 6–7, then dividing these trials into three parcels of 20 trials (first 20 trials of blocks 3–4 and first 20 trials of blocks 6–7 into the first parcel, etc.), and computing *D* scores for each parcel (see Bar-Anan & Nosek, 2014 for similar approach).

Analyses revealed both the image (α = 0.75) and text (α = 0.74) versions of the IAT had similar levels of internal reliability, and a Feldt (1969) test showed that the two measures did not reliably differ in internal reliability $W = 0.992$, $p = .467$.

Both the image and text IAT revealed preferences for gender conforming over androgynous people (Image: $M = 0.21$, $SD = 0.42$, $d = 0.50$; Text: $M = 0.25$, $SD = 0.41$, $d = 0.60$), and comparing between measures found no reliable differences in strength of implicit preferences between groups, $t(807) = 1.21$ $p = .227$, $d = 0.09$, 95% CI = [−0.05, 0.22].

Next, we calculated the correlation between participants' IAT *D* scores and their explicit attitudes, tolerance/intolerance of ambiguity, and beliefs about gender/sex diversity. We found that higher scores on both IATs correlated reliably with greater explicit preferences for male/female over androgynous people, less tolerance of ambiguity, and less affirmative but more normative views of gender/sex diversity (see Table 1 for full correlation matrix and descriptive statistics for image and text IAT). Finally, a series of eight Fisher's *Z* tests revealed no reliable differences between the IATs in correlations with any of the factors except with relation to the gender/sex diversity affirmation scale, where a higher *D* score on the text IAT was more negatively correlated with affirming beliefs about gender/sex diversity than the image IAT (*Z* = 1.98, *p* = .048).

## 5. Study 1b

In Study 1a, both text and image based IATs had acceptable internal reliability and produced moderate to large effects revealing implicit preferences for male/female over androgynous people. Additionally, both versions of the IAT correlated reliably with explicit attitudes about gender expression, intolerance of ambiguity, and some facets of gender/sex diversity beliefs.

To build upon Study 1a, we ran a confirmatory study seeking to replicate those findings using a second image-based IAT with different stimuli. Using another image IAT can ensure that any observed results generalize beyond the particular set of face images used in the first study (see Wells & Windschitl, 1999). To be clear, the use of an image IAT here and in all subsequent studies does not imply a belief that the image IAT is a superior measure to the text IAT, as none of the analyses presented in Study 1a would support that conclusion. Rather, the image IAT was used in an effort to build on parallel work with an image-based IAT measuring implicit attitudes towards transgender people (Axt et al., 2021), and was selected due to being a more compelling measure for our volunteer participant sample.

### 5.1. Method

#### 5.1.1. Participants

A total of 560 volunteers completed the study on the Project Implicit research pool.[3] Data from 21 participants were excluded based on the same criteria used in Study 1a. The final sample of 539 participants (see Appendix A for all demographic information) provided over 80% power to detect a correlation of *r* = 0.12, which was the smallest reliable correlation between the image IAT and an outcome variable in Study 1a. See https://osf.io/ty78e/ for Study 1b's pre-registration.

#### 5.1.2. Measures

##### 5.1.2.1. Implicit attitudes.
Participants completed a second version of the image-based IAT that was identical to the image IAT in Study 1a except for the image stimuli, which consisted of eight faces (four male/female, four androgynous) matched on race and attractiveness. As before, all male/female images were created using headshots taken directly from the Chicago Face Database (Ma et al., 2015), and all androgynous images were created by morphing one of the male/female headshots with another headshot from Chicago Face Database of an individual of the same race and "opposite" gender. Unlike the stimuli used in Study 1a, all images were cropped in an oval shape to include only the front of the face, and exclude all other parts of the head (e.g. hair, ears, etc.), neck, and shoulders (see supplementary materials available at https://osf.io/yvdkx/).

##### 5.1.2.2. Additional measures.
Participants completed the same self-report measures as Study 1a, including explicit attitudes towards gender conformity/androgyny (α = 0.84), the Gender/Sex Diversity Beliefs Scale (α = 0.84) and Tolerance/Intolerance of Ambiguity Scale (α = 0.85).

#### 5.1.3. Procedure
Participants completed all measures in a randomized order.

### 5.2. Results

The image IAT had acceptable levels of internal reliability (α = 0.73) and, replicating Study 1a, *D* scores (*M* = 0.04, *SD* = 0.36) showed more positive implicit associations for gender conforming versus androgynous people on average, t(538) = 2.88 *p* = .004, *d* = 0.12, 95% CI [0.04, 0.21]).

Next, we calculated the strength of correlation between participants' IAT *D* scores and the various self-report measures. Again replicating Study 1a, the image IAT reliably predicted greater explicit preferences for male/female over androgynous people, less tolerance of ambiguity, and less affirmative but more normative and uniform views of gender/sex diversity (see Table 2 for full correlation matrix and descriptive statistics).

### 5.3. Discussion

Results of Study 1b replicated findings from Study 1a for the image IAT, including acceptable levels of internal reliability, reliable implicit preferences for gender conforming people, and correlations with a variety of psychosocial variables. Moreover, the fact that we replicated our findings from Study 1a using a second version of the image IAT suggested that our effects were not a product of our initial stimuli selection. In Study 2, we sought to establish predictive validity between the image version of the IAT and other possible correlates including need for closure, political ideology, and support for nonbinary affirmring policies.

## 6. Study 2

Study 2 tested a larger sample in order to assess incremental predictive validity between level of implicit and explicit gender expression attitudes in predicting political orientation, need for closure, and support for nonbinary affirming policies. While there has been some work showing that more implicit preferences for transgender attitudes are related to support for transgender affirming policies (Axt et al., 2021), no equivalent work has been done examining the relationship between implicit preferences for androgyny and support for nonbinary affirming policies. Moreover, Study 2 used structural equation modeling (SEM) analyses in light of past work showing that linear regression analyses can severely inflate the false positive rate, particularly when using large samples and measures with moderate internal reliability (Westfall & Yarkoni, 2016).

### 6.1. Method

#### 6.1.1. Participants
We originally targeted a sample of 950 participants but delays in study removal allowed for a larger sample. In total, 1118 volunteers completed our study on Project Implicit. Data from 25 participants were excluded from analyses for completing more than 10% of trials faster than 300 milliseconds on the IAT (Nosek et al., 2007). Our final sample of 1093 individuals yielded greater than 95% power to detect a correlation as small as *r* = 0.12. See https://osf.io/86prm/ for Study 2's pre-registration and Appendix A for all demographic information.

---

[3] We initially planned for a sample of 513, but delays in removing the study resulted in slightly more participants.

**Table 1**

Descriptive statistics and correlations for image and text IAT.

| Variable | n | M | SD | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Implicit attitude | 409 | 0.210 | 0.419 | – | 0.274** | 0.070* | −0.172** | −0.363** | 0.188** | 0.095 | 0.184* | 0.070 |
|  | *400* | *0.246* | *0.408* | | | | | | | | | |
| 2. Explicit attitude | 400 | 0.036− | 0.842 | 0.231** | – | 0.197** | −0.109* | −0.501** | 0.359** | 0.349** | 0.198** | 0.068** |
|  | *390* | *0.077* | *0.814* | | | | | | | | | |
| 3. Intolerance of Ambiguity | 399 | 3.853 | 0.790 | 0.127* | 0.210** | – | −0.067 | −0.190* | 0.271** | 0.292** | 0.237** | 0.152** |
|  | *384* | *3.724* | *0.790* | | | | | | | | | |
| 4. Tolerance of Ambiguity | 399 | 5.025 | 0.706 | −0.216** | −0.252** | −0.119** | – | 0.264** | −0.157** | −0.084 | −0.091 | 0.156** |
|  | *389* | *5.072* | *0.671* | | | | | | | | | |
| 5. GSDB[a] Affirmation | 400 | 4.681 | 1.356 | −0.234** | −0.583** | −0.259** | 0.411** | – | −0.564** | −0.313** | −0.416** | −0.135** |
|  | *387* | *4.831* | *1.181* | | | | | | | | | |
| 6. GSDB Gender Normativity | 400 | 1.966 | 0.996 | 0.214** | 0.497** | 0.340** | −0.332** | −0.625** | – | 0.430** | 0.370** | 0.162** |
|  | *386* | *1.848* | *0.872* | | | | | | | | | |
| 7. GSDB Uniformity | 397 | 3.001 | 1.300 | 0.015 | 0.249** | 0.278** | −0.140** | −0.290** | 0.329** | – | 0.126* | 0.150** |
|  | *387* | *2.978* | *0.251* | | | | | | | | | |
| 8. GSDB Surgery | 393 | 2.520 | 1.436 | 0.123* | 0.285** | 0.270** | −0.239** | −0.437** | 0.317** | 0.262** | – | 0.152** |
|  | *383* | *2.499* | *1.446* | | | | | | | | | |
| 9. GSDB Upbringing | 398 | 3.749 | 1.394 | 0.000 | 0.056 | 0.090 | −0.001 | −0.022 | 0.031 | 0.206** | 0.116* | – |
|  | *386* | *3.712* | *1.440* | | | | | | | | | |

[a] GSDB = Gender/Sex Diversity Beliefs Scale; *p < .05, **p < .01. All numbers in non-italic font represent values associated with the Image IAT and all numbers in italics represent values associated with the Text IAT.

**Table 2**

Descriptive statistics and correlations for study 1b.

| Variable | n | M | SD | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Implicit attitude | 539 | 0.044 | 0.353 | – | | | | | | | | |
| 2. Explicit attitude | 534 | −0.065 | 0.783 | 0.203** | — | | | | | | | |
| 3. Intolerance of Ambiguity | 531 | 3.819 | 0.784 | 0.047 | 0.175* | – | | | | | | |
| 4.Tolerance of Ambiguity | 530 | 5.060 | 0.684 | −0.117** | −0.188** | −0.088** | — | | | | | |
| 5. GSDB[a] Affirmation | 527 | 4.809 | 1.263 | −0.210** | −0.478** | −0.254** | 0.323** | – | | | | |
| 6. GSDB Gender Normativity | 525 | 1.909 | 0.958 | 0.161** | 0.363** | 0.328** | −0.196** | −0.543** | — | | | |
| 7. GSDB Uniformity | 527 | 3.176 | 1.348 | 0.141** | 0.227** | 0.326** | −0.030 | −0.260** | 0.336** | — | | |
| 8. GSDB Surgery | 522 | 2.500 | 1.455 | 0.069 | 0.205** | 0.203** | −0.131** | −0.513** | 0.475** | 0.232** | — | |
| 9. GSDB Upbringing | 526 | 3.807 | 1.397 | −0.014 | 0.097* | 0.117** | 0.045 | −0.123** | 0.189** | 0.236** | 0.188** | — |

[a] GSDB = Gender/Sex Diversity Beliefs Scale; *p < .05, **p < .01.

#### 6.1.2. Measures

##### 6.1.2.1. Implicit attitudes.
All participants completed the image-based IAT used in Study 1a. For SEM analyses, the implicit construct was estimated by four indicators, calculated by dividing each IAT block into four bins and creating separate *D* scores for each (see Buttrick et al., 2020).

##### 6.1.2.2. Explicit attitudes.
Participants completed the same explicit preference measure as in Studies 1a-1b. For SEM analyses, the explicit construct was estimated by using standardized responses to the relative preference item and each of the two difference scores concerning positivity and warmth (see Axt et al., 2020 for a similar approach). We again used the average of the three standardized variables to form a single aggregate explicit attitude variable for correlational analyses.

##### 6.1.2.3. Need for closure.
Participants completed the 15-item Need For Closure scale (α = 0.82) to measure the extent to which the desire for cognitive resolution motivates decision making (Kruglanski, 1990; Kruglanski & Webster, 1996; Roets & Van Hiel, 2011). Items were presented in a random order on a 6-point Likert response scale ranging from 1 = *Strongly disagree* to 6 = *Strongly agree*.

##### 6.1.2.4. Support for nonbinary affirming policies.
Participants reported agreement with four nonbinary affirming policies presented in random order on a 7-point Likert response scale ranging from 1 = *Strongly disagree* to 7 = *Strongly agree*. Policies included use of gender-neutral restrooms, changing gender markers on identification documents, gender-neutral parenting, and insurance coverage for gender-affirming care (see supplementary materials available at https://osf.io/yvdkx/ full wording of each item). Higher scores indicated greater support for gender-affirming policies for nonbinary individuals (α = 0.84).

##### 6.1.2.5. Political ideology.
Participants reported their political ideology on a 9-point scale ranging from 1 = *Extremely liberal* to 9 = *Extremely conservative*, with a midpoint of 5 = *Moderate* (Graham, Haidt, & Nosek, 2009).

#### 6.1.3. Procedure
Participants completed the self-report measures in a random order, and completed the image-based IAT either before or after all other measures.

#### 6.2. Results

The image IAT showed acceptable levels of internal reliability (α = 0.75) and again revealed more positive implicit preferences for gender conforming versus androgynous people (M = 0.22, SD = 0.41) on average, t(1092) = 17.22 p < .001, d = 0.52, 95% CI [0.46, 0.58].

Next, we calculated the strength of correlation between participants IAT *D* scores and their explicit attitudes (M = 0.01, SD = 0.86), need for closure (M = 3.79, SD = 0.68), support for nonbinary affirming policies (both individually and in aggregate), and political ideology (M = 4.32, SD = 1.88), and found that higher *D* scores correlated reliably with greater explicit preferences for male/female over androgynous people, more conservative political ideology, less support for nonbinary affirming policies (both individually and in aggregate), and greater need

for closure (see Table 3 for full correlation matrix and descriptive statistics).

Following our pre-registered analysis plan, we then used structural equation modeling to investigate whether or not the IAT predicted meaningful outcomes after accounting for participants' explicit attitudes (e.g. incremental predictive validity). To do this, we constructed a model predicting each outcome (entered as a manifest variable) from both the latent implicit and explicit measures using the indicators detailed in the Measures section.

The latent implicit construct was identified by fixing the path to the first IAT quartering at 1, and the latent explicit construct was identified by fixing the path to the explicit relative preference item at 1. Both the implicit and explicit attitude measures were allowed to covary. This model was compared to a model in which the pathway from implicit measure to outcome measure was constrained to 0. We then used a likelihood ratio test to assess the significance of the difference in fit of the two models, reasoning that incremental predictive validity was demonstrated if there was a significant reduction in model fit when the implicit pathway was removed (see Buttrick et al., 2020).

We found that the IAT showed incremental predictive validity over explicit attitudes for political ideology, and all policy items (both individually and in aggregate), but there was only marginal evidence of incremental predictive validity for need for closure (see Table 4 for SEM output). Stronger implicit preferences for gender conforming over androgynous people was reliably associated with a more conservative political ideology and lower support for policies allowing gender-neutral markers on identification documents, gender-neutral restrooms, gender-neutral parenting, gender-affirming health insurance, and nonbinary affirming policies overall; but only somewhat associated with greater need for closure (see Table 3 for full correlation matrix and descriptive statistics).

### 6.3. Discussion

Study 2 results replicated and extended prior findings, specifically in using SEM analyses to demonstrate incremental predictive validity between implicit attitudes and related beliefs. In particular, stronger implicit preferences for gender conforming over androgynous individuals were significantly associated with a more conservative political ideology, and lower support for policies in favor of gender-neutral markers on identification documents, gender-neutral restrooms, gender-neutral parenting, and gender-affirming health insurance, but not reliably associated with higher need for closure. These results demonstrate the additional value of assessing implicit associations about gender expression above what can be gleaned from examining explicit attitudes alone.

A great deal of prior research seeking to develop novel implicit measures (e.g., Bar-Anan, Nosek, & Vianello, 2009; Bluemke & Friese, 2008; Karpinski & Steinman, 2006; Nosek & Banaji, 2001; Sriram & Greenwald, 2009), or extend such measures into new topics (e.g., Axt et al., 2021; Roddy, Stewart, & Barnes-Holmes, 2010; Rudman & Kilianski, 2000; Banse, Seise, & Zerbes, 2001), have established predictive validity exclusively through self-report. However, given recent discussions of the relationship between measures of implicit associations and relevant behavior (Gawronski, De Houwer, & Sherman, 2020; Kurdi et al., 2019), we sought to add to this literature by conducting three additional tests of the predictive validity of the gender expression IAT on three previously-validated behavioral measures. In addition, we sought to replicate our findings of predictive validity for various self-report measures in another larger sample.

## 7. Studies 3a-c

Studies 3a-c were separate high-powered tests of the gender expression IAT's ability to predict behavior. Specifically, we assessed whether implicit attitudes about androgyny predicted outcomes of other

behavioral measures related to perceptual fluency (Stern & Rule, 2018), judgment bias (Axt, Nguyen, & Nosek, 2018), or mouse-tracking behavior (Hehman, Carpinella, Johnson, Leitner, & Freeman, 2014). Study 3a also included a replication of the test of the IAT's incremental predictive validity concerning various psychosocial variables.

### 7.1. Method

#### 7.1.1. Participants

Participants in all studies came from Project Implicit. In Study 3a, 1502 volunteers completed the study. Data from 39 participants were excluded from analyses using the same criteria as Study 2 resulting in a final eligible sample of 1463 participants (see Appendix A for all demographic information). A minimum of 661 participants were included in each analysis, which allowed for greater than 95% power to detect a correlation as small as $r = 0.14$. See https://osf.io/7stk8/ for Study 3a's registration.[4]

In Study 3b, 1445 volunteers completed the study. Data from 102 participants were excluded from analyses for accepting less than 20% or more than 80% of the applicants in the Judgment Bias Task (or JBT; see Judgment Bias Task in Measures section below), or for accepting or rejecting all of the male/female or androgynous applicants, and data from an additional 21 participants were excluded from analyses using the IAT for having more than 10% of responses faster than 300 ms. A final sample of 1322 eligible participants completed both the gender expression IAT and the JBT, allowing for greater than 95% power to detect a correlation as small as $r = 0.10$. See Appendix A for all demographic information and https://osf.io/dgf7h/ for Study 3b's pre-registration.[5]

In Study 3c, 917 volunteers completed the study (see Appendix A for all demographic information). We excluded data from trials in the mouse tracking task where participants took excessively long to respond (>2000 ms; +3SD from the average response time; Freeman, Pauker, & Sanchez, 2016). These exclusions dropped 10.47% of the responses, including all data from 29 participants. Our final sample of 888[6] eligible individuals allowed 95% power to detect a correlation as small as $r = 0.12$. See https://osf.io/tnfra/ for Study 3c's pre-registration.

### 7.2. Measures

#### 7.2.1. All studies

7.2.1.1. *Explicit and implicit gender expression attitudes.* In all studies, participants completed the same measures of explicit and implicit gender expression attitudes as used in Study 2.

#### 7.2.2. Study 3a

7.2.2.1. *Face evaluation task.* In Study 3a, we assessed the relationship between perceptual fluency (or ease of perception; see Lick & Johnson, 2015) and preferences for gender conforming vs. androgynous faces using the Face Evaluation Task. The task was modeled after a similar paradigm used in Stern and Rule (2018) and consisted of two parts. First,

---

[4] For both studies 3a and 3b, power calculations reflect the expected effect size given a pre-registered sample size of $n = 1300$ for each study. Because our actual sample size was fairly close to the pre-registered number, we report our pre-registered a priori power analysis in the manuscript.

[5] This pre-registration was not formally pre-registered on OSF due to an author's mistake. Thus, the linked document contains the pre-registration information we intended to upload, which was finalized in advance of all data collection for Study 3b in May 2020.

[6] We exceeded our preregistered sample size of $N = 350$ because of a software transition that required the study be taken down at a specific day of the week instead of after achieving a certain number of completed sessions.

**Table 3**

Descriptive statistics and correlations for study 2.

| Variable | n | M | SD | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Implicit attitude | 1093 | 0.219 | 0.420 | – | | | | | | | | |
| 2. Explicit attitude | 1073 | 0.006 | 0.855 | 0.303** | – | | | | | | | |
| 3. Political Ideology | 1062 | 4.320 | 1.884 | 0.260** | 0.447** | – | | | | | | |
| 4. Need for Closure | 993 | 3.790 | 0.678 | 0.131** | 0.229** | −0.187* | – | | | | | |
| 5. Policy 1: Identification[a] | 1072 | 5.130 | 1.802 | −0.253** | −0.478** | −0.554** | −0.167** | – | | | | |
| 6. Policy 2: Restrooms[b] | 1074 | 5.080 | 1.820 | −0.273** | −0.436** | −0.498** | −0.145** | 0.645** | – | | | |
| 7. Policy 3: Parenting[c] | 1067 | 4.010 | 1.928 | −0.239** | −0.438** | −0.479** | −0.154** | 0.504** | 0.500** | – | | |
| 8. Policy 4: Insurance Coverage[d] | 1070 | 4.830 | 1.893 | −0.270** | −0.460** | −0.527** | −0.163** | 0.625** | 0.605** | 0.563** | – | |
| 9. Policy Aggregate[e] | 1077 | 4.765 | 1.533 | −0.311** | −0.550** | −0.622** | −0.188** | 0.836** | 0.830** | 0.784** | 0.849** | – |

[a-d]See supplementary materials available at https://osf.io/yvdkx/ for full-text wording of all policy items; [e]This is an aggregate measure collapsing across all policy items; *$p < .05$, **$p < .01$.

**Table 4**

SEM output table for study 2 and study 3a.

| Outcome | Implicit β | 95% CI | Explicit β | 95% CI | df | -2LL | Δ 2LL | p |
|---|---|---|---|---|---|---|---|---|
| **Study 2** | | | | | | | | |
| Political Ideology | 0.19 | [0.09, 0.29] | 0.62 | [0.53, 0.73] | 8588 | 21,481.1 | 13.59 | <0.001 |
| Policy 1: Identification | −0.17 | [−0.27, −0.07] | −0.67 | [−0.77, −0.57] | 8598 | 21,476.71 | 11.27 | <0.001 |
| Policy 2: Restrooms | −0.22 | [−0.32, −0.12] | −0.57 | [−0.67, −0.48] | 8600 | 21,530.97 | 18.46 | <0.001 |
| Policy 3: Parenting | −0.15 | [−0.25, −0.05] | −0.62 | [−0.72, −0.52] | 8593 | 21,517.44 | 8.5 | 0.004 |
| Policy 4: Insurance Coverage | −0.19 | [−0.29, −0.09] | −0.64 | [−0.75, −0.55] | 8596 | 21,482.6 | 14.63 | <0.001 |
| Policy Aggregate | −0.18 | [−0.27, −0.09] | −0.77 | [−0.87, −0.68] | 8619 | 21,439.03 | 14.47 | <0.001 |
| Need for Closure | 0.10 | [−0.01, 0.22] | 0.31 | [0.21, 0.42] | 8519 | 21,485.56 | 3.17 | 0.075 |
| **Study 3a** | | | | | | | | |
| Political Ideology | 0.23 | [0.10, 0.36] | 0.44 | [0.34, 0.55] | 10,597 | 26,776.7 | 12.7 | <0.001 |
| Policy 1: Identification | −0.22 | [−0.34, −0.10] | −0.59 | [−0.70, −0.49] | 10,601 | 26,715.53 | 12.52 | <0.001 |
| Policy 2: Restrooms | −0.16 | [−0.29, −0.04] | −0.49 | [−0.60, −0.38] | 10,602 | 26,785.74 | 6.53 | 0.011 |
| Policy 3: Parenting | −0.12 | [−0.25, 0.002] | −0.54 | [−0.65, −0.43] | 10,604 | 26,777.87 | 3.72 | 0.054 |
| Policy 4: Insurance Coverage | −0.08 | [−0.21, 0.04] | −0.55 | [−0.66, −0.45] | 10,603 | 26,774.68 | 1.78 | 0.182 |
| Policy Aggregate | −0.18 | [−0.30, −0.06] | −0.66 | [−0.76, −0.56] | 10,607 | 26,700.18 | 8.68 | 0.003 |
| Need for Closure | 0.21 | [0.07, 0.36] | 0.19 | [0.08, 0.31] | 10,535 | 26,682.71 | 8.92 | 0.003 |

participants categorized 48 photos of varying gender phenotype in a random order as either male or female by pressing the E and I keys (see https://osf.io/7wy6s/ for stimuli). Response latency was used as a proxy for ease (or difficulty) of perceptual fluency of the stimuli. Afterwards, participants were shown the same 48 photos one at a time and reported how positive or negative they felt towards each person using a slider response ranging from *0 = Very negative* to *100 = Very positive*. See supplementary materials available at https://osf.io/yvdkx/ for additional details on stimuli creation and to view an example trial from both parts of the task.

*7.2.2.2. Self-report variables.* Participants in Study 3a were randomly assigned to complete either the face evaluation task or the set of self-report variables used in Study 2 consisting of political ideology, opinions about policies related to nonbinary individuals (α = 0.84), and need for closure (α = 0.82).

*7.2.3. Study 3b*

*7.2.3.1. Judgment bias task.* In Study 3b, participants completed a Judgment Bias Task (JBT; Axt et al., 2018). In this version of the JBT, participants were informed that they were evaluating 64 applicants for internships to a new science journalism website (see https://osf.io/9ysed/ for stimuli). They were instructed to accept about half the candidates, and were told to accept the most qualified candidates and reject the least qualified candidates. Participants then passively viewed information about each of the candidates in random order, including each candidate's science and humanities GPAs (each listed on a scale of 1–4), letters of recommendation (rated as poor, fair, good, or excellent), and an interview score from 0 to 100. Participants were instructed to weigh each piece of information equally when evaluating the applicants. Next, participants viewed each applicant's information one at a time

decided whether to either "accept" or "reject" their application.

The JBT was structured so that participants viewed candidates that varied on gender (male or female), gender typicality (androgynous or gender conforming), race (Black, East Asian, Latinx, or White), and qualification (qualified versus unqualified). Determining qualifications followed the same procedure as Axt et al. (2018). Both qualified and unqualified candidates had equal numbers of applicants for each level of gender, gender typicality and race. See supplementary materials available at https://osf.io/yvdkx/ for additional details on stimuli creation.

*7.2.4. Study 3c*

*7.2.4.1. Mouse tracking task.* In Study 3c, participants completed a 120-trial mouse-tracking task. Participants were instructed to sort faces with various gender phenotypes into binary gender categories by moving the mouse from a preset location at bottom-center of the screen to "male" or "female" responses in either top corners (see Fig. 1 for a sample trial; see https://osf.io/3vkx5/ for full set of stimuli). Participants were shown a warning to "move faster" if their mouse failed to move for more than 400 ms during any portion of a trial.

Stimuli in the task consisted of faces that were either 100%, 80%, or 60% male or female. Next, we assigned an "androgyny level" for each face based on the calculation: 100 - |% male - % female|. For example, faces that were 60% female and 40% male (or vice versa) had an "androgyny value" of 80 (100 - |60–40|). This resulted in an androgyny level of 0 (for faces that were 100% male or female), 40 (for faces that were 80% male or female), or 80 (for faces that were 60% male or female). Before the critical tests, participants completed a four-trial training block in which they sorted two faces (not represented in the test stimuli). See supplementary materials available at https://osf.io/yvdkx/ for additional details on stimuli creation.

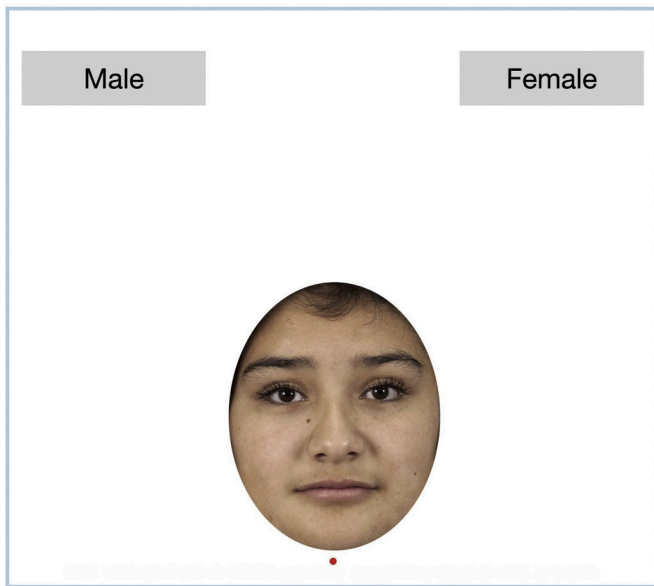In each trial, participants' mouse position on the x-y plane was

**Fig. 1.** Sample trial from mouse tracking task. During each trial, participants saw a face stimulus and were instructed to move their cursor from the red dot (bottom center) to either the "male" or "female" labels in the top corners.

recorded approximately every 15 ms. Based on existing recommendations for mouse tracking analyses (Stillman, Shen, & Ferguson, 2018), we used these continuous location measurements to calculate three specific characteristics; latency, trajectory, and x-flips. Latency was calculated by the amount of time it took to categorize each face, with the assumption that greater latency reflects more decisional conflict (Johnson, Freeman, & Pauker, 2012). Efficiency of mouse trajectory was assessed using area under the curve (or AUC), which was calculated by comparing participants' actual mouse path to an idealized straight trajectory from the starting location to response termination. Larger AUC values (i.e. greater deviations from the straight idealized trajectory) represent greater decisional conflict between response options (Stillman et al., 2018). Finally, uncertainty was calculated through the number of x-flips (i.e., the number of times the mouse reversed direction in the x-plane on each trial). Past work has used x-flips to indicate greater levels of uncertainty and more decisional conflict (Hehman, Stolier, & Freeman, 2015).

### 7.3. Procedure

In Study 3a, participants completed either the face evaluation task or the self-report measures in a random order, completing the IAT as either the first or last measure in the study. In Study 3b, all participants first completed the JBT, followed by the explicit attitudes measure and then the IAT. In Study 3c, participants completed the mouse tracking task followed by the explicit attitude measure and the IAT.

### 7.4. Results

#### 7.4.1. Gender expression IAT

As in prior studies, the image IAT showed acceptable levels of internal reliability (Study 3a $\alpha = 0.75$; Study 3b $\alpha = 0.74$; Study 3c $\alpha = 0.77$), and showed implicit preferences for gender conforming over androgynous people (Study 3a: $M = 0.22$, $SD = 0.41$, $d =$. 0.54; Study 3b: $M = 0.15$, $SD = 0.41$, $d = 0.38$; Study 3c: $M = 0.17$, $SD = 0.42$, $d = 0.40$).

#### 7.4.2. Self-report measures

In Study 3a, greater IAT $D$ scores were again reliably correlated with greater explicit preferences for male/female over androgynous people

($M = 0.01$, $SD = 0.87$), more conservative political ideology ($M = 4.35$, $SD = 2.00$), less support for nonbinary affirming policies (both individually and in aggregate), and greater need for closure (see Appendix K for full correlation matrix and descriptive statistics). Employing the same analysis as in Study 2, we again used structural equation modeling to investigate whether the IAT predicted these outcomes after accounting for participants' explicit attitudes. Here, the IAT showed incremental predictive validity for political ideology, need for closure, policies related to changing gender markers on identification documents, gender-neutral restrooms, and an aggregate policy variable. Effects were marginal for gender neutral parenting and not significant for health insurance coverage for gender-affirming care (see Table 4 for SEM output).

#### 7.4.3. Face evaluation task

Target androgyny was calculated based on the ratio of the morph level associated with each face using the same scoring method as described in Study 3c, resulting in an "androgyny level" of 0 (for faces that were 100% male or female), 40 (for faces that were 80% male or female), or 80 (for faces that were 60% male or female). Then, we used hierarchical linear modeling to predict both reported positivity towards a face and latency in target categorization from target degree of androgyny (Level 1) and participants' explicit and implicit attitudes (Level 2). In the model, latency and positivity for each target were predicted by these variables as well as an interaction between target androgyny and explicit attitudes and an interaction between target androgyny and implicit attitudes. Evidence for predictive validity for the IAT would be seen in the androgyny by implicit attitude interaction term (i.e., whether participants' implicit attitudes moderated the degree to which target androgyny was related to reported positivity towards each target).

For latency (log-transformed), we found a main effect of phenotype such that participants took longer to categorize faces that looked more androgynous ($t(30{,}450.84) = 49.02$, $p < .001$, $SE = 0.001$, $\beta = 0.05$). However, this main effect was not moderated by either explicit ($t(30{,}450.87) = -0.76$, $p = .449$, $SE = 0.01$, $\beta = -0.01$) or implicit ($t(30451) = -1.03$, $p = .302$, $SE = 0.02$, $\beta = -0.02$) attitudes. For liking, there was no main effect of target androgyny ($t(30{,}450.84) = 0.58$, $p = .561$, $SE = 0.25$, $\beta = 0.14$). However, there was an interaction between target androgyny and explicit attitudes ($t(30{,}450.87) = -5.08$, $p < .001$, $SE = 0.28$, $\beta = 1.40$) such that individuals with higher levels of explicit bias against androgynous people had a stronger negative relationship between androgyny and liking. The interaction between implicit attitudes and target androgyny was not significant ($t(30451) = -0.33$, $p = .745$, $SE = 0.58$, $\beta = -0.19$).

#### 7.4.4. Judgment bias task

In Study 3b, the JBT was scored using the same procedure as Axt et al. (2018). Using a Signal Detection Theory framework, a criterion value (operationalized as a measure of strictness or leniency in judgment; see Correll et al., 2007) was calculated separately for gender conforming and androgynous applicants. Lower criterion values indicate more leniency in judgment (i.e., greater likelihood of getting accepted to the science journalism internship regardless of qualifications). A difference score was then created such that higher values on the difference score indicated a lower criterion for gender conforming versus androgynous applicants.

Contrary to expectations, we found an overall negative criterion difference score ($M = -0.10$, $SD = 0.35$, $d = 0.28$), indicating a lower acceptance criterion for androgynous applicants. In other words, participants were more lenient towards androgynous applicants and more strict towards gender typical applicants when deciding whether or accept or reject their applications. In the discussion section, we present some exploratory analyses that suggest possible reasons for why the JBT showed a bias favoring androgynous applicants.

Next, we calculated the strength of correlation between participants'

criterion difference scores and their explicit and implicit attitudes. Here, results showed that criterion difference scores were unrelated to explicit attitudes ($r = -0.002$, $p = .935$) and weakly but negatively related to implicit attitudes ($r = -0.06$, $p = .022$). That is, greater implicit biases against androgynous people were associated with more leniency towards androgynous applicants on the JBT. Given the small and not robust effect, we consider this finding most likely a false positive.

### 7.4.5. Mouse tracking task

We ran a series of three hierarchical linear models separately predicting latency, AUC, and x-flips on each trial by target level of androgyny (Level 1; calculated in the same manner as Study 3a) and implicit and explicit attitudes (Level 2).

Across the three models, we found that target androgyny level predicted latency ($t(80700) = 32.69$, $p < .001$, $SE = 0.0267$, $\beta = 0.873$), AUC ($t(81580) = 10.44$, $p < .001$, $SE = 0.0001$, $\beta = 0.001$), and x-flips ($t(80730) = 14.043$, $p < .001$, $SE = 0.0001$, $\beta = 0.002$), such that higher levels of androgyny predicted increased categorization time, greater deviations from a straight trajectory, and a greater number of x-axis reversals. For interactions between IAT $D$ scores and target androgyny, the only significant analysis was trial latency ($t(80700) = 1.99$, $p = .046$, $SE = 0.06$, $\beta = 0.116$), indicating that the association between facial androgyny and increased latency was stronger among people with stronger implicit preference for gender conforming over androgynous faces. However, IAT $D$ scores did not reliably interact with target androgyny when predicting the other mouse-tracking outcomes (see supplementary materials available at https://osf.io/yvdkx/ for full reporting).

### 7.5. Discussion

Replicating Study 2, SEM analyses in Study 3a found that implicit attitudes about gender expression predicted political ideology, need for closure, and an aggregate measure of gender-related policies after controlling for parallel explicit attitudes. These results again demonstrated the incremental predictive validity of the gender expression IAT beyond parallel explicit attitudes for a number of variables of interest.

However, in a series of exploratory SEM analyses raised during review, we also tested whether implicit attitudes still predicted attitudes about gender-related policies when controlling for both explicit attitudes and political ideology in Study 2 and 3a. Here, we found that both explicit attitudes and ideology consistently predicted policy outcomes individually and in aggregate in both studies (explicit attitudes $\beta$'s range from $[-0.43, -0.30]$ and ideology $\beta$'s range from $[-0.43, -0.31]$ across studies), but implicit attitudes were less consistent, predicting policy outcomes above ideology and explicit attitudes in Study 2 and but not in Study 3a ($\beta$'s range from $[-0.15, -0.09]$ in Study 2 and $[-0.14, 0.003]$ in Study 3a; see Appendix L for full SEM output). Although we found consistent effects for the incremental predictive validity of the gender expression IAT relative to explicit attitudes, results were much less consistent when also accounting for political ideology. Thus, our results do not rule out the possibility that implicit attitudes predict support for gender related policies above and beyond explicit attitudes and political ideology, but suggest that if there is an effect, it is likely very small.

Beyond self-report measures, the gender expression IAT did not reliably predict androgyny-related differences in liking of various faces, treatment of androgynous applicants in a hypothetical judgment task, or mouse tracking behavior in the categorization of targets varying in androgyny. One intriguing and unexpected result from Study 3b was lower criterion (i.e. more leniency) for androgynous over male/female applicants in the JBT. One possible explanation for this effect comes from the treatment of male and female applicants. Prior work using the JBT has occasionally found lower criterion for female versus male applicants (Axt, Casola, & Nosek, 2019), and a similar effect may have impacted Study 3b. Indeed, the criterion value for applicants with 100% male faces ($M = 0.02$, $SD = 0.46$) was reliably higher than for applicants

with 100% female faces ($M = -0.10$, $SD = 0.49$; comparison $d = 0.26$), for applicants with 60% female faces ($M = -0.15$, $SD = 0.46$, comparison $d = 0.38$), and even for applicants with 60% male faces ($M = -0.12$, $SD = 0.47$, comparison $d = 0.31$). Such results suggest that participants may have been recategorizing applicant faces as either "entirely male" or "not entirely male", which would mean that our attempt to match the male/female and androgynous faces on gender was unsuccessful.

Regardless, prior studies have also found evidence of favoritism towards stigmatized groups on the JBT; for example, White participants from a variety of sample sources consistently favored Black over White applicants (Axt, Ebersole, & Nosek, 2016). However, even this "pro-Black" judgment bias was reliably predicted by implicit and explicit racial attitudes. That is, the mere presence of JBT behavior favoring a stigmatized group does not preclude the possibility for such judgments to still be associated with IAT performance.

In our General Discussion, we review several possible reasons for why the gender expression IAT was more predictive of self-report than behavioral measures. However, one possible explanation that appears unlikely is that the measures used in Studies 3a or 3c failed to capture differences in perception or treatment of individuals varying in androgyny. In Study 3a, results found the expected effect of androgyny on latency, with faces with greater operationalized androgyny requiring longer categorization times. Moreover, the relationship between target liking and androgyny was reliably moderated by explicit attitudes, such that people with greater explicit biases against androgynous people in self-report also showed less positivity towards targets with greater androgyny. In addition, Study 3c found the anticipated effects of greater target androgyny creating longer latencies, more x-flips, and more path deviation in mouse-tracking behavior.

## 8. General discussion

When studying social perception, psychology often defaults to studying prototypical identities. As a result, the majority of existing work on gendered face perception has presented binary stimuli (i.e. individuals who appear stereotypically male or female in appearance) and asked participants to make binary judgments (sorting people into categories of male/man or female/woman; Roberts & Bruce, 1988; Brown & Perrett, 1993; O'Toole et al., 1998; Kranz & Ishai, 2006; Dobs, Isik, Pantazis, & Kanwisher, 2019). The failure to include a more diverse representation of gender phenotypes in quantitative and experimental psychology reinforces the status quo assumption that sex assignment at birth always aligns with gender expression, and obscures the rich complexity of gender diversity (Morgenroth & Stuart, 2020). Focusing on binary understandings of sex and gender has implications for research, as it limits both the generalizability of findings and the ability of theories to explain real-world phenomena. This work expands the scope of "who is studied" in research on implicit attitudes to better reflect the diversity of human experience by directly assessing implicit cognition towards androgynous individuals.

### 8.1. Connections to the IAT literature

Our work assessing implicit attitudes about androgyny follows a long tradition of using the IAT as a tool to measure perceptions of stigmatized identities in society, such as race (Greenwald et al., 1998), weight (Schwartz, Chambliss, Brownell, Blair, & Billington, 2003), disability (Pruett & Chan, 2006), age (Rudman, Greenwald, Mellott, & Schwartz, 1999), and sexuality (Sabin, Riskind, & Nosek, 2015). Notably, this work follows the development of a novel Implicit Association Test investigating attitudes towards transgender people (Axt et al., 2021). Participants in our studies demonstrated a robust implicit preference (overall Cohen's $d = 0.44$) for gender conforming over androgynous faces that falls well within the range of prior published effects using the IAT (Greenwald & Lai, 2020). In addition, the aggregate correlation between implicit attitudes and explicit preferences was $r = 0.25$, placing it

squarely in the mid to upper range of prior estimates of implicit-explicit associations (ranging from $r = 0.11$ to $r = 0.36$; (Greenwald & Lai, 2020). Moreover, implicit preferences were associated with several related self-reported measures including intolerance of ambiguity, gender/sex diversity beliefs, need for closure, political ideology, and support for nonbinary affirming policies, and showed incremental predictive validity over participants' explicit attitudes when predicting outcomes like political ideology and policy support. This mirrors previous work from Axt et al., 2021 who also found evidence for incremental predictive validity of the Transgender IAT towards various self-report outcomes like policy support and prior contact with transgender people above and beyond explicit attitudes.

However, implicit attitudes were not reliably linked to any behavioral differences in liking, treatment, or mouse tracking behaviors in response to gender conforming versus androgynous stimuli. There are several possible reasons why we found differences in predictive validity for our self-report versus behavioral measures. One possibility is that self-report variables assessed constructs that were higher in elaboration (Nosek, 2007) than those in our behavioral measures; for example, participants may have already put more thought into self-report measures like support for gender neutral restrooms compared to behavioral measures such as liking of novel faces in Study 3a. To the extent that topics higher in elaboration are associated with greater associations with implicit measures (Nosek, 2005), this discrepancy in elaboration may explain the differences observed in the present work. Another explanation may be more statistical. Specifically, low internal reliability in behavioral measures implies increased measurement error, which could suppress possible associations with the IAT (Kurdi et al., 2019); for instance, our aggregate self-report measure of policy support had noticeably higher internal reliability (median $\alpha = 0.84$) than the liking measure used in Study 3a ($\alpha = 0.60$). Future studies will need to isolate the source of these differences in predictive variability, though we believe the literature benefits from reporting both when implicit attitude measures do and do not predict relevant outcomes (Gawronski, 2019).

### 8.2. Future directions

The IAT used in this work to assess implicit attitudes about androgynous people has a number of potential applications. First, this work assesses attitudes towards people with androgynous *appearance*, which may be distinct from people who *identify* as androgynous. Thus, a logical next step would be to test attitudes towards individuals who self-identify as androgynous to see whether or not the effects from this study generalize to the larger group. Another natural extension is to investigate the implicit attitudes towards androgyny among people who identify as androgynous. Indeed, past research investigating the implicit ingroup attitudes among disadvantaged or stigmatized groups has shown considerable variability (Essien, Calanchini, & Degner, 2020). For instance, while outgroup favoritism in implicit attitudes has been found in domains like age (e.g., Chopik & Giasson, 2017), ingroup favoritism has been found in domains like sexual orientation identity (e. g., Jellison, McConnell, & Gabriel, 2004) or closer to no preference has been shown in domains like race (e.g., Nosek, Banaji, & Greenwald, 2002). Collecting similar data among people who identify as androgynous will aid efforts to both better understand attitudes about gender expression specifically as well as larger theoretical perspectives on implicit ingroup attitudes among members of stigmatized groups.

In addition, recent work suggests that patterns of implicit and explicit attitudes may shift over time. One time-series analysis of data collected on Project Implicit from 2007 to 2016 found that both implicit and explicit attitudes on topics like race, skin-tone, and sexual orientation have become less negative towards stigmatized groups (Charlesworth & Banaji, 2019). In the case of sexual orientation, the authors attribute movement towardtowards neutrality in attitudes to the rapid sociocultural shift in public opinion regarding gay rights and same-sex marriage as well as an increase in opportunities for positive contact

with sexual minorities (Charlesworth & Banaji, 2019; Rosenfeld, 2017). Given this prior research, future work may seek to investigate any long-term changes in attitudes towards androgyny as a possible result of the increased recognition of and exposure to gender diversity in broader society.

Finally, the present work may look to build off recent studies showing successful interventions for changing explicit attitudes towards transgender people. For instance, in-lab experimental manipulations that either presented informational vignettes about transgender identity and images of transgender faces (Flores et al., 2018) or anti-essentialist messages about gender (Wilton et al., 2019) significantly reduced individual levels of transphobia and decreased prejudice towards gender minorities, respectively (Flores et al., 2018; Wilton et al., 2019). Moreover, in a field study in which canvassers knocked on doors and engaged voters in brief perspective-taking conversations about transgender people's rights and experiences led to increased support for a nondiscrimination law protecting transgender people's rights and decreased transphobia that persisted for three months (Broockman & Kalla, 2016). Future work should investigate whether or not these findings also extend to implicit attitudes towards transgender people, and whether these strategies are similarly successful in shifting both implicit and explicit attitudes towards nonbinary people. However, it is worth reiterating that the present results provide no evidence of a causal link between either implicit or explicit attitudes about gender expression and any behavioral outcome measure. As a result, an important first step in such work would be more direct evidence that changes in either explicit or implicit gender expression attitudes mediate changes in outcome variables of interest such as policy support (Forscher et al., 2019).

### 8.3. Limitations

This work is not without shortcomings. First, participants are composed entirely of volunteers from Project Implicit, who are not representative of the broader population in terms of demographic distribution or in their level of motivation to learn about gender biases. For instance, in all six of our studies, females outnumbered all other sex categories at nearly a 2:1 ratio, and our samples were consistently comprised of majority White individuals who are United States citizens residing in the United States. Additionally, we do not have data on participants' current gender identity so are unable to compare responses between individuals who do and do not identify within the gender binary. Finally, it is likely that individuals who volunteer their time and energy participating in online psychology studies do not hold perceptions and attitudes that are representative of the broader human population. Indeed, past research has found that Project Implicit volunteers tend to be younger, more female, and more liberal than US society at large (Charlesworth & Banaji, 2019).

Another potential limitation lies in the artificial nature of the stimuli used to represent androgynous individuals in the IAT, which were created by morphing images from the Chicago Face Database. Although recent work in the context of race found that so long as the faces were not cropped, there were no differences between actual vs artificially morphed multiracial faces in terms of level of attractiveness, distinctiveness, expressivity, or how striking they were (Ma, Kantner, Benitez, & Dunn, 2021), there is also past work suggesting that people may be less sensitive to individual variability in appearance when perceiving artificial, computer generated face stimuli than when perceiving images of real faces (Crookes et al., 2015; Hehman, Sutherland, Flake, & Slepian, 2017). As a result, it is possible that people in our studies perceived the male/female stimuli with greater acuity than the androgynous stimuli. Indeed, some support for this perspective can be seen in data from IAT practice trials, where participants only had to sort androgynous and male/female faces; here, error rates for androgynous faces (9.5%) were slightly higher than error rates for male/female faces (9.1%). Additionally, because our androgynous face stimuli are

artificial, it is unclear how representative they may be of androgynous expression in real life. Finally, we tested a limited number of face stimuli (4 per category across two versions of the IAT for a total of 16 faces; see Appendix B), which limits the generalizability of our work and leaves our findings vulnerable to the influence of stimulus sampling (Wells & Windschitl, 1999).

Finally, recent research suggests that masculinity and femininity may represent two unique dimensions in person perception (as opposed to opposite ends of a single dimension; Hester, Jones, & Hehman, 2020). Thus, our interpretation of androgyny as the midpoint between two ends of a bipolar symmetrical scale may be only one example in a myriad of possible ways that androgyny exists. Other possible representations of androgyny include the simultaneous co-occurrence of high masculinity and femininity (e.g. Bem, 1974), or existing as entirely distinct from masculinity and femininity (Lubinski, Tellegen, & Butcher, 1983). That said, we found moderate correlations between implicit and explicit attitudes about gender expression that are in line with other intergroup domains (Greenwald & Lai, 2020), suggesting that the way androgyny was depicted on the IAT was related to how participants understood androgyny conceptually when reporting their explicit attitudes. Regardless, research on this topic will only benefit from more diverse representations of androgyny.

### 8.4. Conclusion

Motivated by a desire to expand the scope of research in implicit attitudes beyond culturally dominant understandings of sex and gender, we find robust implicit preferences for gender conforming people over androgynous people that correlate reliably with explicit beliefs about gender expression, tolerance of ambiguity, need for closure, and beliefs about gender/sex diversity, and demonstrate that implicit attitudes about gender diversity may predict other outcomes, such as support for specific policies or behaviors towards individuals with varying levels of androgyny. We believe that many other researchers will investigate perceptions of and attitudes towards androgyny, and hope that the measures developed here will facilitate such work. To this end, we have made all data and materials used for this research available at https://osf.io/abruw/. A greater appreciation and understanding of gender diversity will allow psychological research to keep apace with parallel changes in society.

### Open practices statement

All data, study materials, code, and pre-registrations are available at the project page on the Open Science Framework (https://osf.io/abruw/). All measures, manipulations, and exclusions are disclosed. As noted in a footnote on pg. 24 of the manuscript, Study 3b was not formally pre-registered on OSF due to an author's mistake. Thus, the linked document contains the pre-registration information we intended to upload, which was finalized in advance of all data collection for Study 3b in May 2020.

### Declaration of Competing Interest

This research was partly supported by Project Implicit. Dr. Jordan Axt is Director of Data and Methodology for Project Implicit, Inc., a nonprofit organization with the mission to "develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender, or other factors." There are no other potential conflicts of interest with respect to authorship or the publication of this article.

### Acknowledgements

Authors wish to thank Drs. Lindsay Juarez and Nick Buttrick for consulting on analyses, as well as members of the Human Diversity Lab

(at Princeton University), Culture Diversity & Health Lab (at the University of Washington), and two anonymous reviewers for their comments on previous drafts of this paper.

### Appendix A. Supplementary material

Supplementary material for this article can be found online at https://doi.org/10.1016/j.jesp.2021.104162.

### References

Axt, J. R. (2018). The best way to measure explicit racial attitudes is to ask about them. *Social Psychological and Personality Science, 9*(8), 896–906. https://doi.org/10.1177/1948550617728995.

Axt, J. R., Bar-Anan, Y., & Vianello, M. (2020). The relation between evaluation and racial categorization of emotional faces. *Social Psychological and Personality Science, 11*(2), 196–206. https://doi.org/10.1177/1948550619848000.

Axt, J. R., Casola, G., & Nosek, B. A. (2019). Reducing social judgment biases may require identifying the potential source of bias. *Personality and Social Psychology Bulletin, 45*(8), 1232–1251. https://doi.org/10.1177/0146167218814003.

Axt, J. R., Conway, M. A., Westgate, E. C., & Buttrick, N. R. (2021). Implicit transgender attitudes independently predict beliefs about gender and transgender people. *Personality and Social Psychology Bulletin, 47*(2), 257–274. https://doi.org/10.1177/0146167220921065.

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2016). An unintentional, robust, and replicable pro-black bias in social judgment. *Social Cognition, 34*(1), 1–39. https://doi.org/10.1521/soco.2016.34.1.1.

Axt, J. R., Nguyen, H., & Nosek, B. A. (2018). The judgment Bias task: A flexible method for assessing individual differences in social judgment biases. *Journal of Experimental Social Psychology, 76*, 337–355. https://doi.org/10.1016/j.jesp.2018.02.011.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Experimental Psychology, 48*(2), 145–160. https://doi.org/10.1026//0949-3946.48.2.145.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods, 46*(3), 668–688. https://doi.org/10.3758/s13428-013-0410-6.

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology, 56*(5), 329–343. https://doi.org/10.1027/1618-3169.56.5.329.

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General, 147*(8), 1264–1272. https://doi.org/10.1037/xge0000383.

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.

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.

Bluemke, M., & Friese, M. (2008). Reliability and validity of the single-target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology, 38*(6), 977–997. https://doi.org/10.1002/ejsp.487.

Supreme Court of the United States: Bostock v. Clayton County, Georgia. (2020). *Clayton County, Georgia. Certiorari to the United States Court of Appeals for The Eleventh Circuit. No. 17–1618.*

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science, 352*(6282), 220–224. https://doi.org/10.1126/science.aad9713.

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

Buttrick, N., Axt, J., Ebersole, C. R., & Huband, J. (2020). Re-assessing the incremental predictive validity of implicit association tests. *Journal of Experimental Social Psychology, 88*, 103941. https://doi.org/10.1016/j.jesp.2019.103941.

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.

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science, 30*(2), 174–192. https://doi.org/10.1177/0956797618813087.

Chopik, W. J., & Giasson, H. L. (2017). Age differences in explicit and implicit age attitudes across the life span. *The Gerontologist, 57*(suppl_2), S169–S177. https://doi.org/10.1093/geront/gnx058.

Clark, T. C., Lucassen, M. F. G., Bullen, P., Denny, S. J., Fleming, T. M., Robinson, E. M., & Rossen, F. V. (2014). The health and well-being of transgender high school students: Results from the New Zealand adolescent health survey (Youth'12). *Journal of Adolescent Health, 55*(1), 93–99. https://doi.org/10.1016/j.jadohealth.2013.11.008.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology, 92*(6), 1006–1023. https://doi.org/10.1037/0022-3514.92.6.1006.

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

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.

Essien, I., Calanchini, J., & Degner, J. (2020). Moderators of intergroup evaluation in disadvantaged groups: A comprehensive test of predictions from system justification theory. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspi0000302.

Feldt, L. S. (1969). A test of the hypothesis that cronbach's alpha or kuder-Richardson coefficent twenty is the same for two tests. *Psychometrika, 34*(3), 363–373. https://doi.org/10.1007/BF02289364.

Flores, A. R., Haider-Markel, D. P., Lewis, D. C., Miller, P. R., Tadlock, B. L., & Taylor, J. (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.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology, 117*(3), 522–559. https://doi.org/10.1037/pspa0000160.

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., Pauker, K., & Sanchez, D. T. (2016). A perceptual pathway to bias: Interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias. *Psychological Science, 27*(4), 502–517. https://doi.org/10.1177/0956797615627418.

Gawronski, B. (2019). Six lessons for a sogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14*(4), 574–595. https://doi.org/10.1177/1745691619826015.

Gawronski, B., De Houwer, J., & Sherman, J. W. (2020). Twenty-five years of research using implicit measures. *Social Cognition, 38*(Supplement), s1–s25. https://doi.org/10.1521/soco.2020.38.supp.s1.

Geiger, A. W., & Graf, N. (2019, September 5). *About one-in-five U.S. adults know someone who goes by a gender-neutral pronoun*. Pew Research Center. https://pewrsr.ch/2LmF98V.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029–1046. https://doi.org/10.1037/a0015141.

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology, 71*(1), 419–445. https://doi.org/10.1146/annurev-psych-010419-050837.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of Personality and Social Psychology, 85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197.

Harrison, B. F., & Michelson, M. R. (2019). Gender, masculinity threat, and support for transgender rights: An experimental study. *Sex Roles, 80*(1–2), 63–75. https://doi.org/10.1007/s11199-018-0916-6.

Hegarty, P., Ansara, Y. G., & Barker, M. J. (2018). Nonbinary gender identities. In N. K. Dess, J. Marecek, & L. Bell (Eds.), *Gender, sex, and sexualities: Psychological perspectives* (pp. 53–76). Oxford University Press.

Hehman, E., Carpinella, C. M., Johnson, K. L., Leitner, J. B., & Freeman, J. B. (2014). Early processing of gendered facial cues predicts the electoral success of female politicians. *Social Psychological and Personality Science, 5*(7), 815–824. https://doi.org/10.1177/1948550614534701.

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations, 18*(3), 384–401. https://doi.org/10.1177/1368430214538325.

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.

Hester, N., Jones, B. C., & Hehman, E. (2020). Perceived femininity and masculinity contribute independently to facial impressions. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000989.

Howansky, K., Albuja, A., & Cole, S. (2020). Seeing gender: Perceptual representations of transgender individuals. *Social Psychological and Personality Science, 11*(4), 474–482. https://doi.org/10.1177/1948550619875143.

Jellison, W. A., McConnell, A. R., & Gabriel, S. (2004). Implicit and explicit measures of sexual orientation attitudes: In group preferences and related behaviors and beliefs among gay and straight men. *Personality and Social Psychology Bulletin, 30*(5), 629–642. https://doi.org/10.1177/0146167203262076.

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.

Karpinski, A., & Steinman, R. B. (2006). The single category implicit association test as a measure of implicit social cognition. *Journal of Personality and Social Psychology, 91* (1), 16–32. https://doi.org/10.1037/0022-3514.91.1.16.

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.

Kruglanski, A. W. (1990). Motivations for judging and knowing: Implications for causal attribution. In E. T. Higgins, & R. M. Sorrentino (Eds.), *Vol. 2. Handbook of motivation and cognition: Foundations of social behavior* (pp. 333–368). The Guilford Press.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing". *Psychological Review, 103*(2), 263–283. https://doi.org/10.1037/0033-295X.103.2.263.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., … Banaji, M. R. (2019). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *American Psychologist, 74*(5), 569–586. https://doi.org/10.1037/amp0000364.

Leick, G. (2003). *Sex and eroticism in Mesopotamian literature*. Routledge.

Lick, D. J., & Johnson, K. L. (2015). The interpersonal consequences of processing ease: Fluency as a metacognitive foundation for prejudice. *Current Directions in Psychological Science, 24*(2), 143–148. https://doi.org/10.1177/0963721414558116.

Lubinski, D., Tellegen, A., & Butcher, J. N. (1983). Masculinity, femininity, and androgyny viewed and assessed as distinct concepts. *Journal of Personality and Social Psychology, 44*(2), 428–439. https://doi.org/10.1037/0022-3514.44.2.428.

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, D. S., Kantner, J., Benitez, J., & Dunn, S. (2021). Are morphs a valid substitute for real multiracial faces in race categorization research? *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/0146167221989836.

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.

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297.

Morgenroth, T., & Stuart, A. (2020). Gender and psychology. In N. A. Naples (Ed.), *Companion to women's and gender studies* (1st ed., pp. 195–212). Wiley. https://doi.org/10.1002/9781119315063.ch10.

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.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134*(4), 565–584. https://doi.org/10.1037/0096-3445.134.4.565.

Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science, 16*(2), 65–69. https://doi.org/10.1111/j.1467-8721.2007.00477.x.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition, 19* (6), 625–666. https://doi.org/10.1521/soco.19.6.625.20886.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice, 6*(1), 101–115. https://doi.org/10.1037/1089-2699.6.1.101.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The implicit association test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). Psychology Press.

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the implicit association test. *Experimental Psychology, 54*(1), 14–29. https://doi.org/10.1027/1618-3169.54.1.14.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*(1), 36–88. https://doi.org/10.1080/10463280701489053.

Nowak, K. L., & Denes, A. (2016). Androgyny. In A. Wong, M. Wickramasinghe, R. Hoogland, & N. A. Naples (Eds.), *The Wiley Blackwell encyclopedia of gender and sexuality studies* (pp. 1–3). John Wiley & Sons, Ltd.. https://doi.org/10.1002/9781118663219.wbegss282

Osmond, M. W., & Martin, P. Y. (1975). Sex and sexism: A comparison of male and female sex-role attitudes. *Journal of Marriage and the Family, 37*(4), 744. https://doi.org/10.2307/350825.

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.

Perez-Brumer, A., Day, J. K., Russell, S. T., & Hatzenbuehler, M. L. (2017). Prevalence and correlates of suicidal ideation among transgender youth in California: Findings from a representative, population-based sample of high school students. *Journal of the American Academy of Child & Adolescent Psychiatry, 56*(9), 739–746. https://doi.org/10.1016/j.jaac.2017.06.010.

Pruett, S. R., & Chan, F. (2006). The development and psychometric validation of the disability attitude implicit association test. *Rehabilitation Psychology, 51*(3), 202–213. https://doi.org/10.1037/0090-5550.51.3.202.

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.

Rider, G. N., McMorris, B. J., Gower, A. L., Coleman, E., & Eisenberg, M. E. (2018). Health and care utilization of transgender and gender nonconforming youth: A population-based study. *Pediatrics, 141*(3), Article e20171683. https://doi.org/10.1542/peds.2017-1683.

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.

Roddy, S., Stewart, I., & Barnes-Holmes, D. (2010). Anti-fat, pro-slim, or both?: Using two reaction-time based measures to assess implicit attitudes to the slim and overweight. *Journal of Health Psychology, 15*(3), 416–425. https://doi.org/10.1177/1359105309350232.

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.

Romo, V. (2018, July 23). *"Supergirl" casts first transgender superhero on television*. National Public Radio. https://www.npr.org/2018/07/23/631693257/supergirl-casts-first-transgender-superhero-on-television.

Rosenfeld, M. J. (2017). Moving a mountain: The extraordinary trajectory of same-sex marriage approval in the United States. *Socius: Sociological Research for a Dynamic World, 3*. https://doi.org/10.1177/2378023117727658.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the implicit association test. *Social Cognition, 17*(4), 437–465. https://doi.org/10.1521/soco.1999.17.4.437.

Rudman, L. A., & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and Social Psychology Bulletin, 26*(11), 1315–1328. https://doi.org/10.1177/0146167200263001.

Sabin, J. A., Riskind, R. G., & Nosek, B. A. (2015). Health care providers' implicit and explicit attitudes toward lesbian women and gay men. *American Journal of Public Health, 105*(9), 1831–1841. https://doi.org/10.2105/AJPH.2015.302631.

Schmidt, S. (2019, September 17). *Merriam-Webster adds non-binary pronoun "they" to dictionary*. The Washington Post. https://www.washingtonpost.com/dc-md-va/2019/09/17/merriam-webster-adds-non-binary-prounoun-they-dictionary/.

Schudson, Z. C., & van Anders, S. M. (2021). Gender/sex diversity beliefs: Scale construction, validation, and links to prejudice. *Group Processes & Intergroup Relations*. https://doi.org/10.1177/1368430220987595.

Schwartz, M. B., Chambliss, H. O., Brownell, K. D., Blair, S. N., & Billington, C. (2003). Weight bias among health professionals specializing in obesity. *Obesity Research, 11*(9), 1033–1039. https://doi.org/10.1038/oby.2003.142.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology, 56*(4), 283–294. https://doi.org/10.1027/1618-3169.56.4.283.

Budner, S. (1962). Intolerance of ambiguity as a personality variable. *Journal of Personality, 30*(1), 29–50. https://doi.org/10.1111/j.1467-6494.1962.tb02303.x.

Stern, C., & Rule, N. O. (2018). Physical androgyny and categorization difficulty shape political conservatives' attitudes toward transgender people. *Social Psychological and Personality Science, 9*(1), 24–31. https://doi.org/10.1177/1948550617703172.

Stillman, P. E., Shen, X., & Ferguson, M. J. (2018). How mouse-tracking can advance social cognitive theory. *Trends in Cognitive Sciences, 22*(6), 531–543. https://doi.org/10.1016/j.tics.2018.03.012.

Surtees, A., & Dyer, J. (2020). *Exploring gender diversity in the ancient world (intersectionality in classical antiquity)* (1st ed.). Edinburgh University Press.

Swim, J. K., & Cohen, L. L. (1997). Overt, covert, and subtle sexism: A comparison between the attitudes toward women and modern sexism scales. *Psychology of Women Quarterly, 21*(1), 103–118. https://doi.org/10.1111/j.1471-6402.1997.tb00103.x.

Valdes, F. (1994). *Queers, Sissies, Dykes, and Tomboys: Deconstructing the conflation of" sex," "gender," and "sexual orientation" in Euro-American law and society [Doctoral Dissertation]*. Stanford University.

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.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS One, 11*(3), Article e0152719. https://doi.org/10.1371/journal.pone.0152719.

Wilson, B. D. M., Choi, S. K., Herman, J. L., Becker, T., & Conron, K. J. (2017). *Characteristics and mental health of gender nonconforming adolescents in California: Findings from the 2015–2016 California health interview survey*. The Williams Institute and UCLA Center for Health Policy Research. http://healthpolicy.ucla.edu/publications/Documents/PDF/2017/gncadolescents-factsheet-dec2017.pdf.

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.

Wittlin, N. M., Dovidio, J. F., LaFrance, M., & Burke, S. E. (2018). About face: Memory for transgender versus cisgender targets' facial appearance. *Journal of Experimental Social Psychology, 78*, 77–92. https://doi.org/10.1016/j.jesp.2018.04.009.