Journal of Experimental Psychology: Human Perception and Performance

Determinants of Shared and Idiosyncratic Contributions to Judgments of Faces

Daniel N. Albohn, Joel E. Martinez, and Alexander Todorov

Online First Publication, September 19, 2024. https://dx.doi.org/10.1037/xhp0001239

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology:
Human Perception and Performance

ISSN: 0096-1523

https://doi.org/10.1037/xhp0001239

Determinants of Shared and Idiosyncratic Contributions to Judgments of Faces

Daniel N. Albohn[1], Joel E. Martinez[2], and Alexander Todorov[1]
[1] Booth School of Business, The University of Chicago
[2] Department of Psychology, Harvard University

Recent work has shown that the idiosyncrasies of the observer can contribute more to the variance of social judgments of faces than the features of the faces. However, it is unclear what conditions determine the relative contributions of shared and idiosyncratic variance. Here, we examine two conditions: type of judgment and diversity of face stimuli. First, we show that for simpler, directly observable judgments that are consistent across observers (e.g., masculinity) shared exceeds idiosyncratic variance, whereas for more complex and less directly observable judgments (e.g., trustworthiness), idiosyncratic exceeds shared variance. Second, we show that judgments of more diverse face images increase the amount of shared variance. Finally, using machine-learning methods, we examine how stimulus (e.g., incidental emotion resemblance, skin luminosity) and observer variables (e.g., race, age) contribute to shared and idiosyncratic variance of judgments. Overall, our results indicate that an observer's age is the most consistent and best predictor of idiosyncratic variance contributions to face judgments measured in the current research.

Public Significance Statement

Group-level models of judgment have provided important insights into how individuals view others. However, emerging evidence suggests that group-level averages explain only a portion of the reliable variance, necessitating models that represent the idiosyncrasies of the individual. Despite this evidence, little is known about the relative contributions of the stimulus and perceiver. In the current work, we provide evidence that the variance explained by the stimulus and perceiver is not fixed, but rather depends on theoretically and practically important factors, including the type of judgment being examined and the diversity of the stimulus set being evaluated.

Keywords: social judgments, variance partitioning, idiosyncratic models

Supplemental materials: https://doi.org/10.1037/xhp0001239.supp

The dominant approach for studying how individuals form judgments about others has been to draw conclusions based on aggregated metrics collected across many different observers. This approach is suitable for studying variations in judgments that result from stimulus differences that influence judgment systematically and consistently across observers, such as a neutral face incidentally resembling an emotional expression (Oosterhof & Todorov, 2008; Zebrowitz, 2017). These systematic stimulus differences are a source of shared variance, reflected in consensual judgments across observers.

However, the dominant approach fails to capture variations in judgments that result from perceivers' differences, such as emotional disposition, age, gender, political ideology, and other personality traits—all of which have been shown to influence how people interpret faces (Albright et al., 1988; Ebner, 2008; Kenny & La Voie, 1984; Malloy et al., 2023; Mattarozzi et al., 2015). This variation also results from interactions of the stimulus and perceiver. That is, meaningful variance depends on the individual perceiving specific stimuli, such as one individual weighting masculine facial features more predominantly than others in their judgments of attractiveness or an individual with depression interpreting a neutral face as threatening (Leppänen et al., 2004).

While there is evidence showing consensus in judgments about a particular stimulus (e.g., a face) across observers, when and why certain individuals' judgments are and are not similar to the group average has been an area understudied within the judgment and decision-

Isabel Gauthier served as action editor.

The authors have no known conflicts of interest to disclose. This study's design and hypotheses were preregistered (https://aspredicted.org/9he9e.pdf). Data and code related to this article are posted online (https://osf.io/4scj5/).

Daniel N. Albohn served as lead for data curation, formal analysis, investigation, methodology, project administration, software, validation, visualization, and writing–original draft and served in a supporting role for conceptualization. Joel E. Martinez served in a supporting role for writing–original draft. Alexander Todorov served as lead for funding acquisition, resources, and supervision, contributed equally to investigation, project administration, validation, and visualization, and served in a supporting role for formal analysis and writing–original draft. Daniel N. Albohn, Joel E. Martinez, and Alexander Todorov contributed equally to writing–review and editing. Joel E. Martinez and Alexander Todorov contributed equally to methodology and conceptualization.

Correspondence concerning this article should be addressed to Daniel N. Albohn, Booth School of Business, The University of Chicago, 5807 South Woodlawn Avenue, Chicago, IL 60637, United States. Email: Daniel.Albohn@chicagobooth.edu

making literature. Work that has explicitly examined these factors has been largely descriptive and has merely documented that idiosyncrasies exist in human judgment. However, understanding what factors influence when judgments are shared between individuals and when they are highly idiosyncratic has both theoretical (e.g., which groups are likely to share judgments) and practical (e.g., study design) implications. To this end, the present work attempts to move beyond descriptive analyses and toward explaining conditions under which idiosyncratic judgment contributions may be larger than shared judgment contributions, or vice versa, and what stimulus- and participant-level features might predict such sources of variance.

Shared and Idiosyncratic Contributions

The typical judgment study rarely partitions the sources of variance in judgments, perhaps because such partitioning requires repeated judgments of the stimuli (Martinez et al., 2020). The reported traditional measures of reliability on aggregated data essentially mask individual differences. For example, Cronbach's α is partially confounded with test length. A lengthy judgment task with many judges could result in a large reliability score despite low interrater reliability (Hönekopp, 2006; Todorov, 2017, Chapter 7). Some studies do attempt to partition the variance between stimuli and individuals, but often ignore the critical interaction between participant and stimulus (e.g., Hehman et al., 2017). As such, these methods treat a substantial amount of meaningful variance as noise, but understanding this "noise" variance is critical for building comprehensive models of perception and judgment (Kahneman et al., 2021; Martinez et al., 2020).

A principled approach is to explicitly model individual-, stimulus-level, and interaction influences on judgments (Hönekopp, 2006; Martinez et al., 2020). The stimulus-level, shared contribution is an estimate of the effect of a particular stimulus on judgments. For example, a smiling face might be judged higher on trustworthiness across individuals compared to a neutral face. In this case, stimulus features explain a large portion of the variance observed in judgments. On the other hand, participant-level, idiosyncratic contributions are an estimate of an individual's judgment either across all stimuli judged (i.e., the main effect of participant) or relative to other stimuli and other participant judgments (i.e., the interaction between participant and stimulus; Hönekopp, 2006; Martinez et al., 2020).

The "main effect" participant variance is less universally interpreted as an "idiosyncratic" contribution as differences in the variance could be due either to differences in preferences between individuals or differences in how each individual interprets the judgment scale (Hönekopp, 2006; Martinez et al., 2020). This variance component will be large if there are large differences between the mean judgments of participants, even though the relative ranking of stimuli might be the same across participants. On the other hand, the participant-by-stimulus interaction variance component is less ambiguous in its interpretation as it relies on differences in the relative ranking of the stimuli. For example, Rater A might find Stimulus 1 more attractive than Stimulus 2, but Rater B might show the opposite pattern.

Determinants of Shared and Idiosyncratic Contributions

Within the framework outlined above, a handful of studies have estimated the contribution of shared and individual variance in face judgments. In the case of attractiveness judgments, for example, upwards of 50% of the variance is explained by idiosyncratic preferences and less than 30% is explained by shared preferences (Hehman et al., 2017; Hönekopp, 2006; Martinez et al., 2020). In contrast, judgments such as femininity, masculinity, and age consistently yield higher shared relative to idiosyncratic variance contributions (Albohn et al., 2022; Bjornsdottir et al., 2022; Hehman et al., 2017). These studies suggest that the type of judgment is one of the most important determinants of the relative contributions of shared and idiosyncratic variance. Correspondingly, in all studies, we examined how the type of judgment influences variance estimates. We predicted that judgments with a clear mapping to physical cues—that is, a mapping that is consistent between observers (e.g., masculinity, femininity)—would show greater shared relative to idiosyncratic contributions. In contrast, judgments with a mapping that is less consistent between observers (e.g., perceived trustworthiness) would show greater idiosyncratic relative to shared contributions.

In Study 2, we also examined how the diversity of face stimuli influences shared and idiosyncratic contributions to judgments. In the very few studies examining this question, the evidence has been mixed. Hehman et al. (2017) reported that less emotionally intense faces (e.g., more neutral in appearance) and real, naturally varying faces showed greater idiosyncratic variance compared to emotionally intense faces and computer-generated standardized faces. Presumably, both lack of expressivity and extraneous image features (background, clothes, posture, etc.) leave room for different individual interpretations of the stimuli, resulting in larger idiosyncratic variance. On the other hand, Hönekopp (2006) noted that highly homogeneous stimuli may also increase idiosyncratic contributions to variance. To illustrate this point, imagine judging the attractiveness of the faces of fashion models as opposed to a representative sample of the population. Although the former would be judged as more attractive, it is possible that the relative ranking of faces would be more idiosyncratic in the restricted range of attractiveness. Faces that are highly standardized—such as those collected in the lab—are likely to have a similarly restricted range. Therefore, we predicted that judgments of more diverse faces should result in increased shared and decreased idiosyncratic variance relative to judgments of more homogenous faces. Finally, we explored how additional factors–such as the type of scale used to measure judgments–might influence the variance partitioning in a Supplemental Study (see the online supplemental materials).

Predictors of Shared and Idiosyncratic Variance

As noted above, considerable progress has been achieved in understanding the stimulus features that influence judgments of faces (e.g., Oosterhof & Todorov, 2008; Stirrat & Perrett, 2010; Zebrowitz, 2017). However, identifying predictors of idiosyncratic variance has been exceedingly difficult. Factors that one might think should exert an influence on judgments, such as culture or genetics, provide very little explanatory power in individual preferences (Germine et al., 2015; Hester et al., 2021; Sutherland et al., 2020). Instead, the single consistent factor predictive of individual preferences is personal environment, which is, by definition, made up of many features (Germine et al., 2015; Sutherland et al., 2020). Despite the lack of direct evidence showcasing important predictors of idiosyncratic variance, past work has shown that individuals weigh features differently when making judgments. For example, some individuals find highly

masculine faces attractive while others find them unattractive (Oh et al., 2020; Said & Todorov, 2011; Zietsch et al., 2015). Similarly, individual differences in self-reported extraversion are predictive of facial masculinity preferences (Welling et al., 2009). However, the individual factors related to the differential weighting of face features are much less understood outside of the attractiveness literature.

In Study 3, we explored what stimulus- and participant-level factors predict the shared and idiosyncratic contributions to judgments. We examined whether stimulus-level features such as incidental resemblance to emotion expressions, skin tone, and facial features assumed to be related to women and men (i.e., gender-stereotype dimorphism) could explain the shared contribution of judgments and, more importantly, whether stimulus- and participant-level features such as race, gender, and age could explain the idiosyncratic contribution to judgments.

Study 1: Type of Judgment

The objective of Study 1 was to examine whether the amount of shared relative to idiosyncratic variance was dependent on the type of judgment being evaluated. Specifically, we sought to examine the variance of judgments with a clear mapping to physical cues that is consistent across observers (perceived masculinity and femininity) and the variance of judgments with less consistent mapping across observers (perceived trustworthiness and attractiveness). The selection of these judgments was informed by a pilot study (see the online supplemental materials), wherein participants judged the "inferability/observability" of a number of judgments from faces. Whereas "masculinity" and "femininity" were judged as highly inferable and observable from the face, "trustworthiness" was judged as difficult to infer and observe. Although "attractiveness" was judged as highly inferable and observable, prior work shows that most of the meaningful variance is idiosyncratic (Hönekopp, 2006; Martinez et al., 2020). Because of the unique nature of this judgment and its long history of study (Rhodes, 2006; Rhodes et al., 1998), it was also included.

Method

Transparency and Openness

All data and scripts are available online (https://osf.io/4scj5/). Each study's design and analyses were preregistered (https://aspredicted.org/9he9e.pdf) and all deviations from the preregistration are reported. All studies reported in this manuscript were reviewed and approved by The University of Chicago Institutional Review Board. Likewise, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. Data for all studies were collected in 2023.

Participants

One hundred and ninety-nine participants ($M_{age} = 39.99$, $SD_{age} = 10.97$) were recruited through CloudResearch. Participants self-identified as follows: 97 women, 101 men, one trans (transgender, trans man, trans women); nine Asian, 24 Black, seven Latinx, one Middle Eastern or North African, 157 White, and one other/not reported. Participants were located within the United States and "CloudResearch approved" (i.e., participants prescreened by CloudResearch to meet certain quality standards; Hauser et al., 2023).

We randomly assigned each participant to a single judgment condition, aiming for approximately 40 participants per judgment.[1] Sample size was determined by following guidelines established by Martinez et al. (2020) through simulations, with a recommended minimum of 40 participants per judgment task given the number of stimuli used in all experiments ($N_{stimuli} = 120$). Due to complete randomization, we ended up with the following number of participants per judgment: 56 for "attractive," 48 for "feminine," 45 for "masculine," and 50 for "trustworthy."

Stimuli

Stimuli were 120 neutral face images from the Chicago Face Database (Ma et al., 2015) that were selected to be representative in terms of both sex/gender and race/ethnicity (the same stimuli used in Jones et al., 2021). The self-disclosed demographics of the selected images were: 30 Black (15 male; 15 female), 30 White (15 male; 15 female), 30 Asian (15 male; 15 female), and 30 Latinx faces (15 male; 15 female). The average perceived age of these stimuli was 26.38 based on norming data (Ma et al., 2015).

Experimental Procedures

This study was run online. After consenting, participants were taken to a screener page designed to filter nonhuman participants. Participants were then instructed that they would be rating 120 individuals on a specific personality judgment two times and that they should not take too long on any single judgment. Next, participants were randomly assigned to a judgment condition and were shown the 120 images in a random order with the prompt, "How [JUDGMENT] is this individual?" (i.e., attractive, trustworthy, masculine, or feminine). Participants responded on a 7-point Likert-type scale, ranging from "1" = *not at all* to "7" = *very*, using the number keys on their keyboard.

After participants rated all 120 faces, they were instructed that they would continue to the second part of the experiment (rerating the images) after a self-paced break. When participants decided to continue, they again rated each of the same 120 stimuli in a new random order on the same judgment assigned to them.

After both rating blocks were completed, participants filled out a short demographic questionnaire (race, ethnicity, gender, and age), two debriefing questions ("What did you think this study was about?" and "Did you notice anything odd or off about the images?"), were told the purpose of the study, and compensated for their time ($2.50 USD; $M_{time} = 13.94$ min).

Analytic Procedures

We partitioned the variance for each judgment into shared variance contributions (stimulus-level) and idiosyncratic variance contributions (both participant-level variance and participant-by-stimulus-level variance) following procedures outlined by Martinez et al. (2020). Specifically, we constructed a series of linear mixed-effects regressions with random intercepts using the lme4 package for R (Bates et al.,

[1] The preregistration stated our goal was 50 participants per condition (i.e., four judgment conditions × 50 participants per judgment = 200 total) in an effort to ensure at least 40 participants per condition, which was determined to be an appropriate sample size per condition given the number of trials we included; for details, see simulation data in Martinez et al. (2020).

2015; R Core Team, 2023). We included random intercepts for each participant, stimulus, and block (the first or second time the stimulus was rated), as well as random intercepts for participant-by-stimulus, participant-by-face, and participant-by-block interactions. We used the "bobyqa" optimization in lme4 in order to minimize convergence issues (Bates et al., 2015).

The focus of the presented results is the variance partition coefficient (VPC). This metric represents the proportion of variance explained by the cluster of interest in relation to the total amount of variance. It is formally defined as,

$$\mathrm{VPC} = \frac{\sigma^2_{\mathrm{cluster}}}{\sigma^2_{\mathrm{cluster}} + \sigma^2_{\mathrm{residual}}}. \quad (1)$$

As noted in Martinez et al. (2020),

> when the VPC is closer to one, the between-cluster variance will explain most of the variance and the within-cluster variance will be low. When the VPC is closer to zero, there is no between-cluster variance, and thus most of the variance is within-cluster. (p. 1558)[2]

Although the VPC analysis is the main focus of the current paper, these metrics are largely descriptive and do not allow for statistical tests between studies (e.g., how does the diversity of stimuli affect the amount of shared variance?). To supplement this analysis, we computed several correlations: (a) the intrarater correlation as a measure of test–retest reliability, which also acts as an upper bound limit of variance explained for each participant; (b) the interrater correlation, which indicates the average agreement at the level of individual participants; and (c) the correlation of each participant's judgments to the group average judgments (calculated without that participant's data). The latter correlation is the most straightforward to interpret as a measure of shared preference, with larger correlations suggesting high shared preferences and smaller correlations suggesting lower shared preferences. We conduct statistical tests on these correlations (after converting them to Fisher's *z*-scores) to measure both the effect of the type of judgment and the diversity of face stimuli. We report analyses on the relevant correlations across studies after we report each study's variance partitioning results.

Results and Summary

Examining the VPCs revealed that most of the meaningful variance for femininity and masculinity judgments was explained by shared contributions. As displayed in Figure 1, these contributions accounted for approximately 54% of the observed variance in femininity judgments, whereas idiosyncratic contributions accounted for a combined total of 31% of the variance (i.e., participant plus participant-by-stimulus interaction). Likewise, shared contributions accounted for approximately 47% of the variance in masculinity judgments, whereas idiosyncratic contributions accounted for a combined total of 32% of the variance. The observed variance of the main effect of the rating block, as well as its interaction with the participant and stimulus, was small (<1% for both femininity and masculinity).

While shared contributions explained most of the variance in judgments of femininity and masculinity, this was not the case for judgments of attractiveness and trustworthiness. Idiosyncratic contributions for attractiveness accounted for nearly 60% of the variance in the model, whereas stimulus-level, shared contributions accounted for approximately 10% of the variance. Likewise, idiosyncratic contributions for trustworthiness judgments accounted for a combined total of 52% of the variance, whereas shared contributions accounted for approximately 4.2% of the observed variance. As before, the observed variance of the main effect of the rating block, as well as its interaction with the participant and stimulus, was nominal for both judgments (1.2% and 2%, respectively).

Further, and as predicted, the shared, idiosyncratic, and reliability correlation coefficients were all larger for judgments of femininity and masculinity compared to judgments of attractiveness and trustworthiness (see Table 1).

As predicted, the relative contributions of idiosyncratic and shared variance to preferences varied as a function of the type of judgment. For judgments with a clear mapping to physical cues that is consistent across observers (i.e., femininity and masculinity), the variance accounted for by the stimulus-level, shared component was higher than the variance accounted for by both idiosyncratic components. In contrast, for judgments with a mapping that is not as consistent across observers (i.e., attractiveness and trustworthiness), the idiosyncratic variance was larger than the shared variance. Taken together, the findings show that the type of judgment had a large influence on the amount of variance accounted for in each cluster of interest. In particular, people tended to agree on whether someone appeared feminine or masculine because such judgments are likely derived from physical features that individuals consistently agree on (e.g., a square jaw is stereotypically masculine). However, people's judgments of attractiveness and trustworthiness were largely idiosyncratic, most likely because the same physical cues are interpreted and weighted differently by different people (e.g., one individual may have a preference for round faces, while another individual may have a preference for angular faces).

Study 2: Type of Stimulus

Study 2 was designed to examine whether the amount of shared relative to idiosyncratic variance was influenced by the heterogeneity of the stimulus faces. Study 1 (and the Supplemental Study in the online supplemental materials) used highly standardized face images (i.e., similar background, lighting, angle, etc.). In contrast, Study 2 used images randomly selected from a large database of neutral faces that varied naturally. These stimuli are less standardized, more diverse (in appearance and demographics), and less homogenous in terms of cues for specific judgments (e.g., attractiveness).

Method

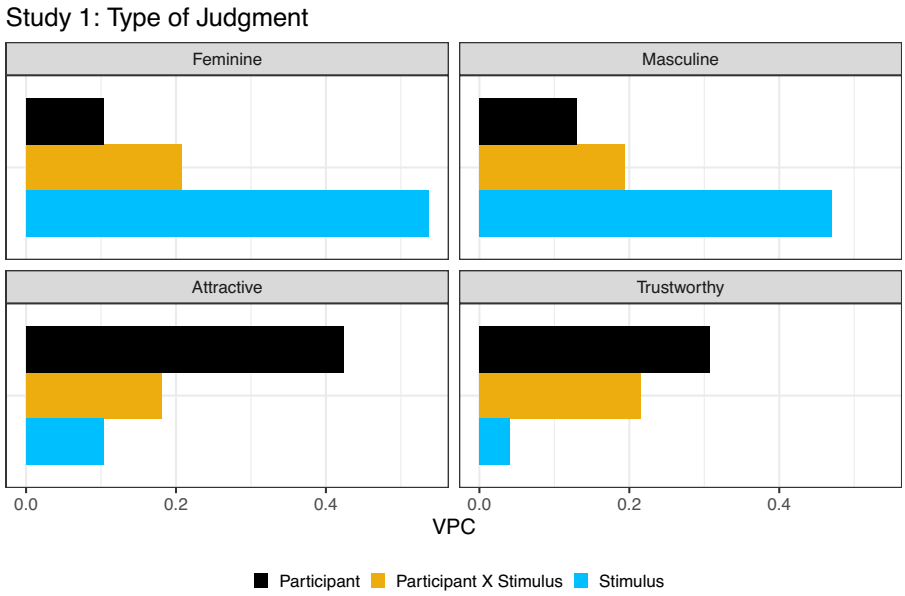
Participants

One hundred and ninety-nine participants ($M_{\mathrm{age}} = 40.62$, $SD_{\mathrm{age}} = 12.89$) were recruited through CloudResearch (one participant's data failed to collect). Participants self-identified as follows: 75 women, 119 men, three nonbinary, and two other/not reported; 23 Asian, 18 Black, 11 Latinx, 143 White, and four other/not reported. Participants were located within the United States and "CloudResearch approved."

[2] As per our preregistration, we also computed beholder indices (Hönekopp, 2006) which are reported in the online supplemental materials.

Figure 1

VPC for Each Measured Judgment in Study 1



Note. The VPC scores (*x* axis) represent the amount of variance explained by each grouping factor (black bars for participant, orange bars [dark grey] for participant X stimulus, and blue bars [light grey] for stimulus varaince). The stimulus variance factor represents "shared" variance, or the amount of variance that is shared and consistent between individual participants. The participant and participant-by-stimulus factors represent the two forms of "idiosyncratic" variance, or the amount of variance that is attributable to the individual participant. The participant variance component can be interpreted as differences in mean stimulus judgments or differences in scale interpretation. The participant-by-stimulus interaction can be interpreted as participants judging stimuli differently. VPC = variance partition coefficients. See the online article for the color version of this figure.

As with Study 1, we randomly assigned each participant to a single judgment, aiming for approximately 40 participants per judgment. Due to complete randomization, we ended up with the following number of participants per judgment: 55 for "attractive," 51 for "feminine," 46 for "masculine," and 47 for "trustworthy."

Stimuli

Stimuli were randomly selected from a larger, internal pool of neutral face images collected from various databases for machine learning (no standardization) and Internet image searches. Briefly, the entire pool of images underwent a screening process in which each image was first classified as a real, neutral face by both an expert reviewer and at least one pretrained emotion detection algorithm (e.g., Amazon's Rekognition or DeepFace). The database was further cleaned by removing any faces that were too small once aligned ($<$1,000 $\times$ 1,000 px), blurry or of low quality, severely oblique facing, or of well-known individuals (e.g., famous politicians or celebrities). Face images were center aligned and cropped to 1,024 $\times$ 1,024 pixels. This yielded a final database of approximately 8,800 high-quality, aligned neutral face images. Next, the pretrained facial attribute classifier, DeepFace, was used to estimate face race, gender, and age (Serengil & Ozpinar, 2021). This model was trained using FairFace, a face dataset that is balanced in terms of perceived race and gender and includes norming data for perceived race, gender, and age (Karkkainen & Joo, 2021). We used the classifications provided by DeepFace to randomly select stimuli from the larger pool of faces that reflected the same number and rated demographic makeup as the previous study: 30 Black (15 male; 15 female), 30 White (15 male; 15 female), 30 Asian (15 male; 15 female), and 30 Latinx faces (15 male; 15 female) all within the perceived estimated age range of 18–50 years old.

Procedure

Study procedures were identical to Study 1 with the exception of the stimuli described above ($M_{\text{time}} = 14.17$ min).

Results and Summary

Replicating Study 1 but with different stimuli (Figure 2), shared contributions accounted for 73% of the variance, whereas idiosyncratic contributions accounted for a combined total of 17% of the variance in femininity judgments. The variance for the main effect of the rating block, as well as its interaction with the participant and stimulus, was less than 1%. Similarly, shared contributions accounted for 60% of the variance in masculinity judgments, whereas idiosyncratic contributions accounted for a combined total of 28% of the variance. The variance of the main effect of the rating block, as well as its interaction with the participant and stimulus, was less than 1%.

Table 1
Correlation Coefficients and Reliability Estimates With 95% Confidence Intervals for Each Judgment Examined in Studies 1 and 2

Judgment	Correlation	Study 1	Study 2
Feminine	Leave one out average (shared preference)	.77 [0.71, 0.82]	.90 [0.87, 0.92]
	Interrater average (idiosyncratic preference)	.60 [0.50, 0.68]	.81 [0.76, 0.87]
	Reliability (intrarater)	.74 [0.67, 0.80]	.88 [0.84, 0.91]
	Cronbach's α average	.99 [0.99, 0.99]	1.0 [1.0, 1.0]
Masculine	Leave one out average (shared preference)	.75 [0.69, 0.81]	.81 [0.77, 0.85]
	Interrater average (idiosyncratic preference)	.57 [0.46, 0.66]	.66 [0.58, 0.72]
	Reliability (intrarater)	.69 [0.61, 0.76]	.83 [0.78, 0.87]
	Cronbach's α average	.99 [0.99, 0.99]	.99 [0.99, 0.99]
Attractiveness	Leave one out average (shared preference)	.49 [0.36, 0.61]	.68 [0.58, 0.76]
	Interrater average (idiosyncratic preference)	.25 [0.09, 0.41]	.47 [0.33, 0.59]
	Reliability (intrarater)	.52 [0.40, 0.63]	.71 [0.56, 0.75]
	Cronbach's α	.95 [0.94, 0.96]	.98 [0.98, 0.98]
Trustworthiness	Leave one out average (shared preference)	.26 [0.10, 0.42]	.39 [0.23, 0.52]
	Interrater average (idiosyncratic preference)	.08 [−0.10, 0.25]	.17 [−0.01, 0.33]
	Reliability (intrarater)	.34 [0.18, 0.47]	.66 [0.56, 0.75]
	Cronbach's α average	.83 [0.78, 0.87]	.88 [0.84, 0.90]

Note. Leave-one-out correlation (shared preference) is the average correlation between each participant's responses and the mean group responses (without that participant's data); the interrater correlation is the average correlation between each participant's responses with every other participant's responses; the intrarater correlation is the average of every participant's test–retest correlation between rating Blocks 1 and 2. Cronbach's α reports the internal consistency among faces and is best interpreted as the expected correlation between the aggregated judgment and the aggregated judgment of another sample (with the same size) of participants.

On the other hand, attractiveness judgments were primarily explained by idiosyncratic contributions, accounting for nearly 47% of the variance. Stimulus-level shared contributions accounted for 31% of the variance. The variance of the main effect of the rating block, as well as its interaction with the participant and stimulus, was less than 1%. Similarly, idiosyncratic contributions for trustworthiness judgments accounted for a combined total of 69% of the variance, whereas shared contributions accounted for approximately 8% of the variance. The variance of the main effect of the rating block, as well as its interaction with the participant and stimulus, was small (1.3%).

Finally, as shown in Table 1, all correlation metrics were large and increased in value from Study 1, suggesting that there was more shared agreement between participants in Study 2 compared to Study 1.

Comparing the variance estimates of Study 1 and Study 2 (Figure 2), there was more variance explained by shared, stimulus-level features when the stimuli varied naturally compared to when they were highly standardized. Both the shared VPC and shared preference correlation coefficients increased across all judgments in Study 2 compared to Study 1. The idiosyncratic VPC for the participant-by-stimulus interaction showed a similar pattern, with three out of the four increasing in Study 2 or comparable to Study 1. This pattern of results was accompanied by the VPC main effect of the rater decreasing in Study 2 across all judgments compared to Study 1. In summary, Study 2 showed that increasing the diversity of the stimulus set used to solicit judgments also increased the amount of variance attributable to shared, stimulus-level features.

Shared Preference Comparisons Across Studies

While VPCs are informative, they are largely descriptive and cannot be statistically compared across studies. The shared preference correlation coefficient metric provides a straightforward and interpretable value for each participant. Therefore, we can compare the amount of reliable variance captured by shared preference between judgments and across studies at the individual level. As a reminder, all individual correlation coefficients were first Fisher transformed to z values; p values are adjusted for multiple comparisons where necessary.

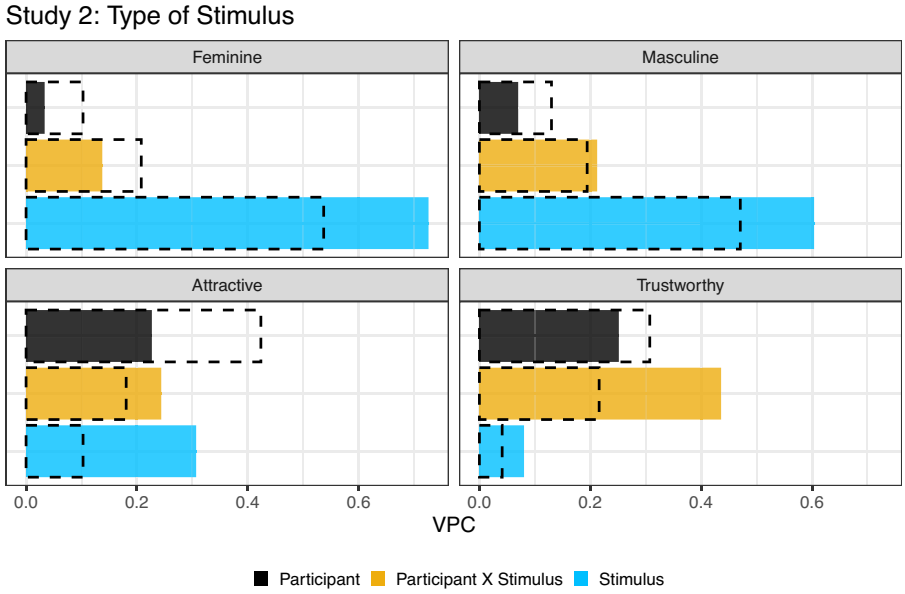
Study 1 Correlation Analysis: The Influence of Judgment Type

We first examined whether the shared correlation coefficient differed between each of the four judgments (attractiveness, trustworthiness, masculinity, or femininity). There was a significant effect of judgment, $F(3, 196) = 57.62$, $p < .001$, $\eta^2 = .47$. Pairwise comparisons revealed that all judgments significantly differed from one another aside from judgments of femininity and masculinity, $t(91) = 0.84$, $p = .403$, $d = 0.17$. Importantly, the correlation was significantly higher for judgments of femininity and masculinity than judgments of attractiveness and trustworthiness: femininity versus attractiveness judgments, $t(103) = 7.84$, $p < .001$, $d = 1.54$; and versus trustworthiness judgments, $t(96) = 10.80$, $p < .001$, $d = 2.18$; masculinity versus attractiveness judgments, $t(100) = 7.18$, $p < .001$, $d = 1.43$; and versus trustworthiness judgments, $t(93) = 10.52$, $p < .001$, $d = 2.16$. Finally, the correlation was higher for attractiveness than trustworthiness judgments, $t(105) = 5.60$, $p < .001$, $d = 1.09$.

Study 1 Compared to Study 2 Correlations: The Influence of Stimulus

The comparisons of the shared preference correlation coefficients between judgments in Study 2 were nearly identical to those of Study 1, $F(3, 195) = 74.23$, $p < .001$, $\eta^2 = .53$ (see the online supplemental materials for all pairwise comparisons).

Figure 2
VPC for Each Judgment in Study 2



Note. The VPC scores (*x* axis) represent the amount of variance explained by each grouping factor (black bars for participant, orange bars [dark grey] for participant X stimulus, and blue bars [light grey] for stimulus varaince). Dashed lines represent VPC values from Study 1 (see Figure 1 for more details). The stimulus variance factor represents "shared" variance, or the amount of variance that is shared and consistent between individual participants. The participant and participant-by-stimulus factors represent the two forms of "idiosyncratic" variance, or the amount of variance that is attributable to the individual participant. The participant variance component can be interpreted as differences in mean stimulus judgments or differences in scale interpretation. The participant-by-stimulus interaction can be interpreted as participants judging stimuli differently. VPC = variance partition coefficients. See the online article for the color version of this figure.

We also compared the shared preference correlation coefficients between the two studies to test whether using nonstandardized stimuli significantly increased the amount of agreement. The two main effects of judgment (attractiveness, trustworthiness, masculinity, or femininity), $F(3, 390) = 130.55$, $p < .001$, $\eta^2 = .50$, and study (Study 1 vs. Study 2), $F(1, 390) = 34.45$, $p < .001$, $\eta^2 = .08$, were significant, but their interaction was not significant, $F(3, 390) = 1.12$, $p = .34$, $\eta^2 = .01$.

Summary

Overall, our correlation analyses replicated the patterns observed through the VPC analysis. In particular, femininity and masculinity judgments had significantly more shared preference than attractiveness and trustworthiness judgments across all studies (including our Supplemental Study in the online supplemental materials examining how response scale influenced variance contributions). This pattern reflects the VPC analyses, which showed higher shared variance for judgments of femininity and masculinity.

Our correlation metric of shared preference significantly increased in Study 2 from Study 1 across all judgments measured, underscoring that the ratio of shared-to-idiosyncratic variance is not fixed. In this particular case, less standardization in the stimuli being evaluated resulted in more agreement between participants, ostensibly due to the characteristics of the stimuli themselves. Critically, however, the same correlation pattern across judgments emerged irrespective of the stimuli (Study 2) or response scale (Supplemental Study in the online supplemental materials): judgments with a clear, consistent mapping to physical features, in particular masculinity and femininity, resulted in larger shared correlations.

Study 3: Predictors of Shared and Idiosyncratic Variance

The objective of Study 3 was to determine stimulus- and participant-level features that are significant predictors of the specific variance components calculated in the previous studies. To accomplish this, we used XGBoost, a powerful machine-learning model (Chen & Guestrin, 2016). We chose XGBoost because it is an interpretable model when used in conjunction with Shapley additive explanations (commonly referred to as SHAP and based on Shapley values), a method that quantifies important features in the model (Lundberg & Lee, 2017).

Motivation and Preliminary Investigation

Across all studies (Studies 1 and 2 and the Supplemental Study in the online supplemental materials), we find clear and large differences in the amount of shared and idiosyncratic variance between judgments of masculinity and femininity and judgments of attractiveness and trustworthiness. One potential source of these differences is the weighting of facial features when making these types of judgments. For judgments with a clear mapping to physical features, such as masculinity and femininity, participants'

weighting should be consistent. For example, almost all participants should rate faces with larger eyes as more feminine and less masculine (Diego-Mas et al., 2020; Paunonen et al., 1999). For more complex judgments, such as attractiveness and trustworthiness, participants' weighting need not be consistent (e.g., while some participants rate faces with large eyes as trustworthy, others rate them as untrustworthy). As an illustration of this differential weighting, we plotted the correlation densities between each participant's judgments of faces and two measured face features: eye size and face roundness (see Study 3 Method for details on how each face feature was measured).

Specifically, we correlated each participant's ($N_{participant} = 398$) face judgments with the faces' measurements of eye size and roundness ($N_{face} = 120$). The correlation density plots for masculinity ($N_{participant} = 91$) and femininity ($N_{participant} = 99$) judgments with eye size and face roundness are as expected (see Figure 3). For the majority of participants, eye size and face roundness were positively correlated with judgments of femininity and negatively correlated with judgments of masculinity. On the other hand, the correlations between attractiveness judgments ($N_{participant} = 111$) and eye size and face roundness were positive for many participants, indicating that these participants agreed that larger eyes and rounder faces were more attractive. But there was a nontrivial number of participants who had negative correlations, suggesting that these participants found larger eyes and rounder faces less attractive. Finally, the correlation between judgments of trustworthiness ($N_{participant} = 97$) and eye size and face roundness were approximately normally distributed around 0, indicating a high degree of idiosyncrasy in how participants weighted eye size and face roundness in their judgments of trust worthiness.

This example indicates that individuals do utilize facial features differently, particularly when making judgments that are highly idiosyncratic, such as attractiveness and trustworthiness. Informed by this initial, illustrative example, we tested whether we could identify stimulus- and participant-level features that are predictive of shared and idiosyncratic variance.

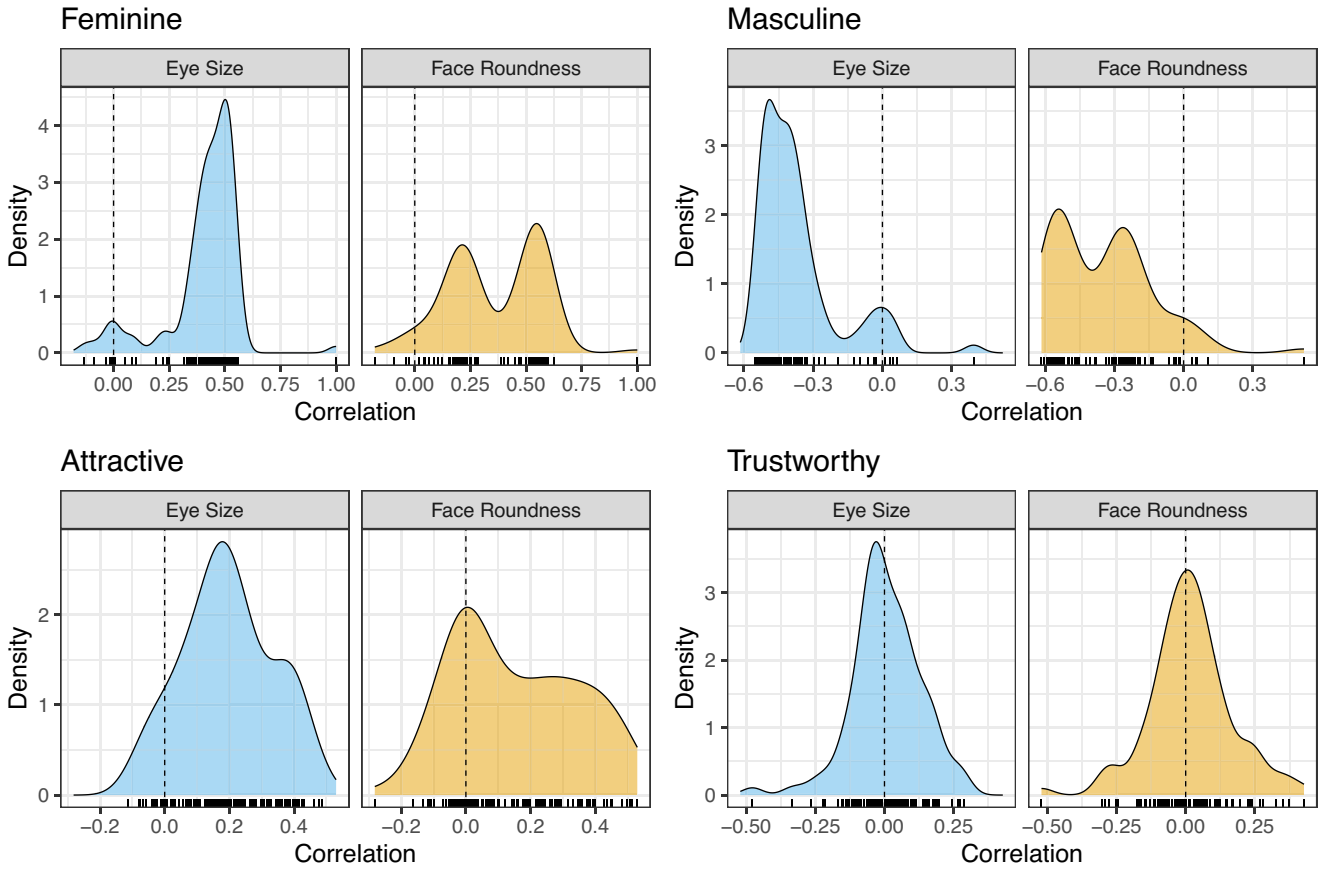
Method

Data

We combined the data from Studies 1 and 2 in order to boost the amount of training data, as the studies were equivalent in everything aside from the type of stimuli that were evaluated by participants.

Figure 3

Densities of Correlations Between Participants' Judgments (Masculine, Feminine, Attractive, and Trustworthy) of Faces and Measured Face Features (Eye Size and Face Roundness)



Note. Eye size and face roundness were measured independently of judgments through pixel-level analysis using automated facial landmark detection (dlib; King, 2009); the measures were standardized to account for variation between images. See the online article for the color version of this figure.

Thus, we re-calculated our linear mixed-effects regression model using this combined data set. Next, we estimated the stimulus and participant-by-stimulus random effect estimates from the regression model using 5,000 posterior resampling simulations to ensure we obtained stable coefficients. These coefficients represent the portion of the response unique to the stimulus or participant (relative to the overall mean). The median values of the resampled coefficients were used as the dependent variables in our predictive models.

Training Features

Participant-level training features were derived from self-report measures of participant race, gender, and age (i.e., the demographic information they filled out). The stimulus-level features of race, gender, and age were estimated using the categorizations provided by DeepFace (see Study 2 Method for details). Stimulus-level face metrics (e.g., eye size, face height) and incidental emotion resemblance (i.e., the likelihood that the neutral face resembled an angry, happy, sad, disgusted, fearful, or surprised expression) were estimated using a modified and updated version of a neutral face-specific stacked-ensemble deep learning model developed and detailed in previous studies (for details see, Adams et al., 2022; Albohn & Adams, 2021). Additionally, we measured stimulus face width, face height, cheekbone prominence, eye size, and mean interior face luminance (skin only) through pixel-level analyses (e.g., the distance in pixels of each face's width at its widest point; King, 2009). Pixel-level measurements were standardized by dividing each metric by its face's eye distance to control for differences across images, such as camera angle or lens distortion.

To predict the shared judgment coefficients, we only used the stimulus-level features; to predict the idiosyncratic judgment coefficients, we used both the participant- and stimulus-level features as our idiosyncratic coefficients represented the participant-by-stimulus interaction coefficients from our models.

Accuracy and Bias in Algorithms

It is important to note that while we use pretrained algorithms to categorize some aspects of our training and test faces, such models should not be considered ground truth (Todorov et al., 2022). That is to say, an algorithm may determine that a face is most likely a 43-year-old Asian woman when in fact any or all of those classifications may be correct or incorrect to varying degrees. The models we used were trained using responses from human participants and thus inherit all of the biases and complications that occur when averaging participant responses. The output from these models "projects" its best estimate onto the face, just like when experimenters use the average race, gender, or age classification response from several hundred participants to determine the "most likely" agreed upon category for a particular stimulus. In both cases, the determined classifications need not align with a certain racialized criterion (e.g., the portrayed participant's self-disclosed racialized identity) and thus can result in discordance between self-understanding and perceived classification. However, we believe that (a) we have taken careful consideration to use language that reflects "perceived" rather than accurate classifications, and (b) that the average, perceived social categories provided by the model can lend important insight into what factors are important for predicting shared and idiosyncratic contributions to judgments. For example, self-disclosed racialized identity may be less informative than perceived race for predicting variance. Ideally, both self-disclosed and average perceived social demographic features of the stimuli being judged would be examined in the model to more comprehensively account for various racialization processes (see Martinez, 2023; Roth, 2016). However, in the present study, we were limited by the nature of the stimulus sets used and thus focused primarily on perceived attributes of the stimuli.

Results

Analysis Strategy

The predictor variables outlined above were entered into the appropriate XGBoost model (either a "shared" model with stimulus-specific predictors, or an "idiosyncratic" model with all predictors) after the following feature preprocessing: numeric predictor variables were normalized, categorical variables were dummy-coded, and predictors with no variance were removed. After an initial model was fit, we performed basic hyperparameter tuning using 800 model iterations with eight hyperparameters that were free to vary (tree depth, number of trees, learning rate, number of parameters to try at each split, minimum number of data points in a node, loss reduction required for split, number of iterations with improvement before stopping, and sample size of data exposed to each iteration). The tuning was determined through Latin hypercube sampling. All models were fit using fivefold cross-validation and were evaluated on the full dataset during hyperparameter tuning (i.e., to identify the best model parameters for our data). After hyperparameter tuning, we refit and evaluated the best model using 100 different training and test 80/20 data splits to obtain unbiased performance metrics. Table 2 reports the average training and test root-mean-square error and R^2 values for each variance condition and judgment.

Next, we interrogated each model's predictive power by computing the SHAP values for each predictor in the model. Each SHAP value can be positive or negative and adding their sum to the model's baseline (i.e., mean predicted value) results in the predicted value from the model; these values are best interpreted as how much a certain feature contributes to the model's predicted value for a given observation. We focus on a global interpretation of the strength of

Table 2
Summary Statistics for the Best XGBoost Model Predicting the Shared or Idiosyncratic Linear Mixed-Effects Regression Coefficients Across Each Judgment Examined

	Shared: Stimulus		Idiosyncratic: Participant × Stimulus	
Judgment	RMSE	R^2	RMSE	R^2
Feminine	(.74, .97)	(.85, .73)	(57, .73)	(.60, .19)
Masculine	(.54, .84)	(.90, .73)	(.57, .73)	(.61, .24)
Attractive	(.39, .59)	(.79, .46)	(.64, .64)	(.05, .02)
Trustworthy	(.26, .36)	(.85, .08)	(.62, .74)	(.41, .02)

Note. Each numeric pair in parentheses corresponds to the average training and testing metric, respectively (e.g., [train, test]), for models fit on 100 random 80/20 train/test data splits. RMSE = root-mean-square error.

each predictor, which corresponds to the mean absolute value of the SHAP score (see the online supplemental materials for additional feature importance results). The SHAP values are on the same scale as the dependent variable, which for the presented models is the median shared (stimulus-level; $M = 0$, $SD = 1.31$) or idiosyncratic (participant-by-stimulus; $M = 0$, $SD = 0.76$) coefficient from our linear mixed-effects regression models.

It should be noted that the SHAP values represent the important features that the model uses for prediction, not important features that human participants use for judgments, per se. However, all of the predictive features in each model were directly (e.g., participant gender, age) or indirectly (e.g., emotion similarity, perceived gender) calculated from human responses or were physical measurements directly perceptible to human judges (e.g., face width, luminance). Likewise, if important model predictors align with predictors that past research would suggest are important to human perception, we can reasonably conclude that the predictors are important across both machine and human judges (e.g., face width related to judgments of femininity or masculinity). That said, the purpose of examining important model predictors is not necessarily to investigate a one-to-one correspondence between a machine model and a human judge, but rather to determine when and under what circumstances features become important for examining the idiosyncrasies of judgment. For example, one might intuit that participant sex/gender is an important predictor of judgments of femininity or masculinity, yet if this information yields little predictive power to the model, it should be noted for future research.

Shared Variance Predictors

As displayed in Figure 4, the most important predictors for shared preference coefficients are in line with what one would expect for a given judgment. Perceived stimulus gender was most important for predictions related to attractiveness ($|M| = 0.18$), femininity ($|M| = 1.16$), and masculinity ($|M| = 0.85$). Likewise, other cues related to perceived gender-stereotype dimorphism such as eye size, face width, and face luminance were also important features, albeit significantly less compared to the perceived gender of the face. On the other hand, the most important stimulus features for predicting shared preference coefficients of trustworthiness judgments were the likelihood that the stimulus appeared like a happy expression and not like a sad expression, followed by several perceived gender-stereotype cues (i.e., eye size, face roundness).

Idiosyncratic Variance Predictors

The important model features for predicting idiosyncratic preference coefficients appeared to be largely stimulus-driven rather than features related to the demographics of the participant, with the exception of participant age (see Figure 5). Interestingly, participant age was the most important feature for prediction across all four judgments measured ($|M_{\text{attractive}}| = 0.01$, $|M_{\text{trustworthy}}| = 0.04$, $|M_{\text{femininity}}| = 0.10$, $|M_{\text{masculinity}}| = 0.11$). While participant race appeared influential for predicting judgments of trustworthiness ($|M| = 0.02$) and stimulus gender was important for predicting femininity ($|M| = 0.08$) and masculinity ($|M| = 0.08$) judgments, these features appear to be second compared to the importance of participant age, underscoring that other participant demographic features contribute little to the overall prediction in these idiosyncratic models.

Summary

Across both the shared and idiosyncratic models, important predictive features were generally aligned with what past research suggests is important for human judges. For example, stimulus features such as perceived gender and eye size were important predictors for shared variance across all judgments. Interestingly, stimulus properties were also identified as important features in the idiosyncratic models. In contrast, the self-reported participant demographics (aside from participant age) were relatively unimportant to idiosyncratic variance predictions.

Comparing feature importance across the shared and idiosyncratic models revealed two important insights. First, the shared variance models performed much better than the idiosyncratic models. Second, highly idiosyncratic judgments are harder to predict (even for shared models), as evidenced by smaller feature importance values compared to highly shared judgments, such as femininity and masculinity.

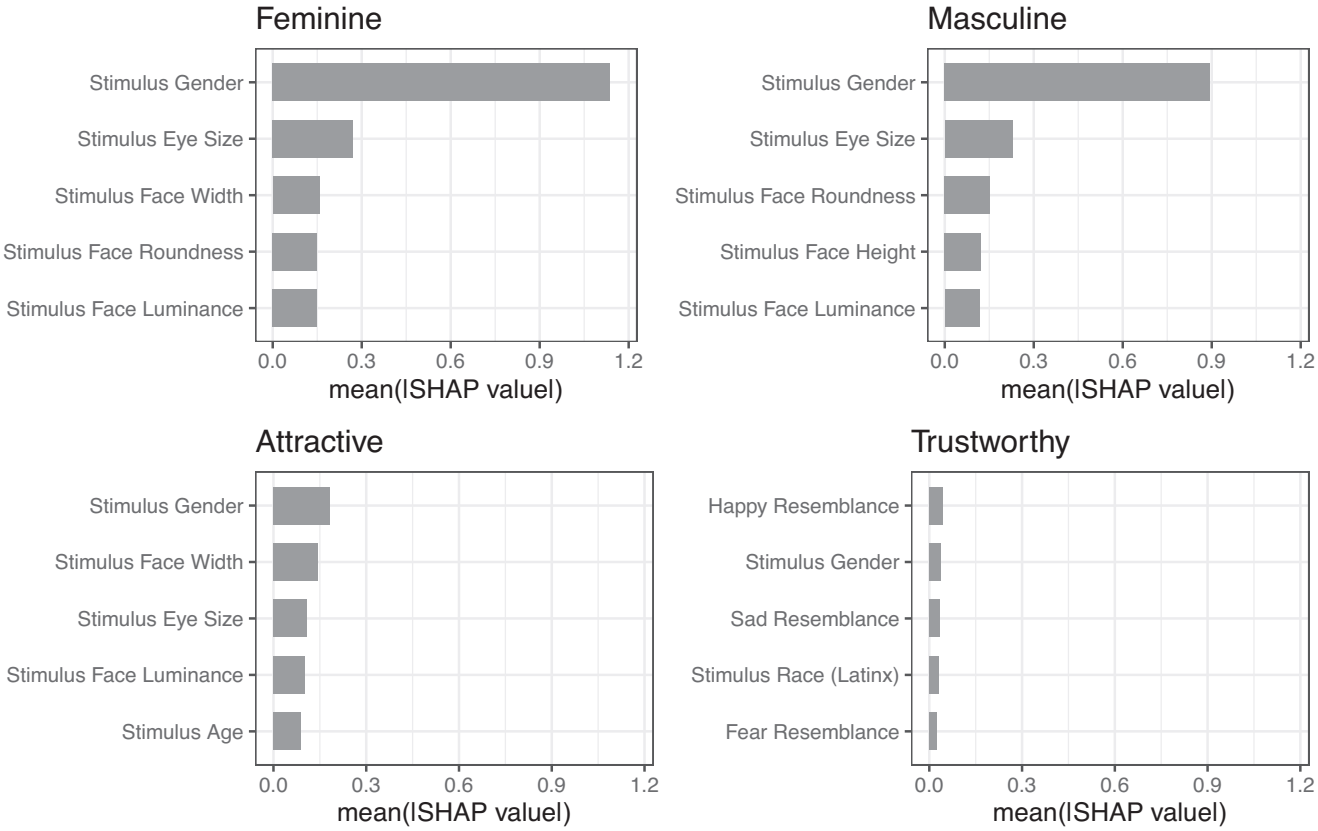
General Discussion

The way in which individuals derive information from faces is a critical component of a holistic understanding of person perception. Prior work within this domain has largely focused on how individuals make judgments of others on average. That is, responses and analyses are aggregated across all individuals sampled and inferences are drawn about judgments as a group-level behavior. However, a growing body of literature suggests that a substantial portion of the variance of judgments is not explained by models that aggregate judgments (Albohn et al., 2022, 2024; Hönekopp, 2006; Kurosu & Todorov, 2017; Martinez et al., 2020; Peterson et al., 2022; Todorov & Oh, 2021; Zhan et al., 2021). While this unexplained variance is traditionally treated as "noise," a sizable portion is also meaningfully related to the idiosyncrasies of the participants making judgments; oftentimes over half of the meaningful variance is explained by participant idiosyncrasies. Given the growing importance of idiosyncrasies in perception and judgment, and the wide application that this research has across disciplines, it is paramount for researchers to understand when and under what circumstances participant idiosyncrasies will dominate judgments over shared, stimulus-level features.

The goals of the present research were twofold. First, we sought to examine what factors might influence the ratio of shared to idiosyncratic contributions to judgments derived from faces. Across two reported studies and an additional supplemental study, we found that (a) the type of judgment being measured, (b) the sample of faces used to solicit judgments, and (c) the scale used to measure judgments (see Supplemental Study in the online supplemental materials) all influenced this ratio. By far the largest factor that increased shared relative to idiosyncratic variance contributions was the type of judgment, a pattern that remained irrespective of the diversity of faces or scale used. Judgments with a clear mapping to physical face features—that is consistent across observers—contain more shared than idiosyncratic variance. In contrast, judgments with a mapping that is less consistent across observers contain more idiosyncratic than shared variance. The diversity of stimuli also influenced the shared-to-idiosyncratic variance contribution ratio, albeit less than the type of judgment. Specifically, judgments of more diverse stimuli increased the shared variance relative to

Figure 4

Mean Absolute SHAP Values (x Axis) for the Top Five Most Important Features of Each Shared Variance Model



Note. The SHAP value of each feature can be interpreted as the degree to which that feature, on average, affects the final predicted result of the model. Values are additive, meaning that each SHAP value adds that amount to the model's final predicted value relative to the model's average prediction. SHAP = Shapley additive explanation.

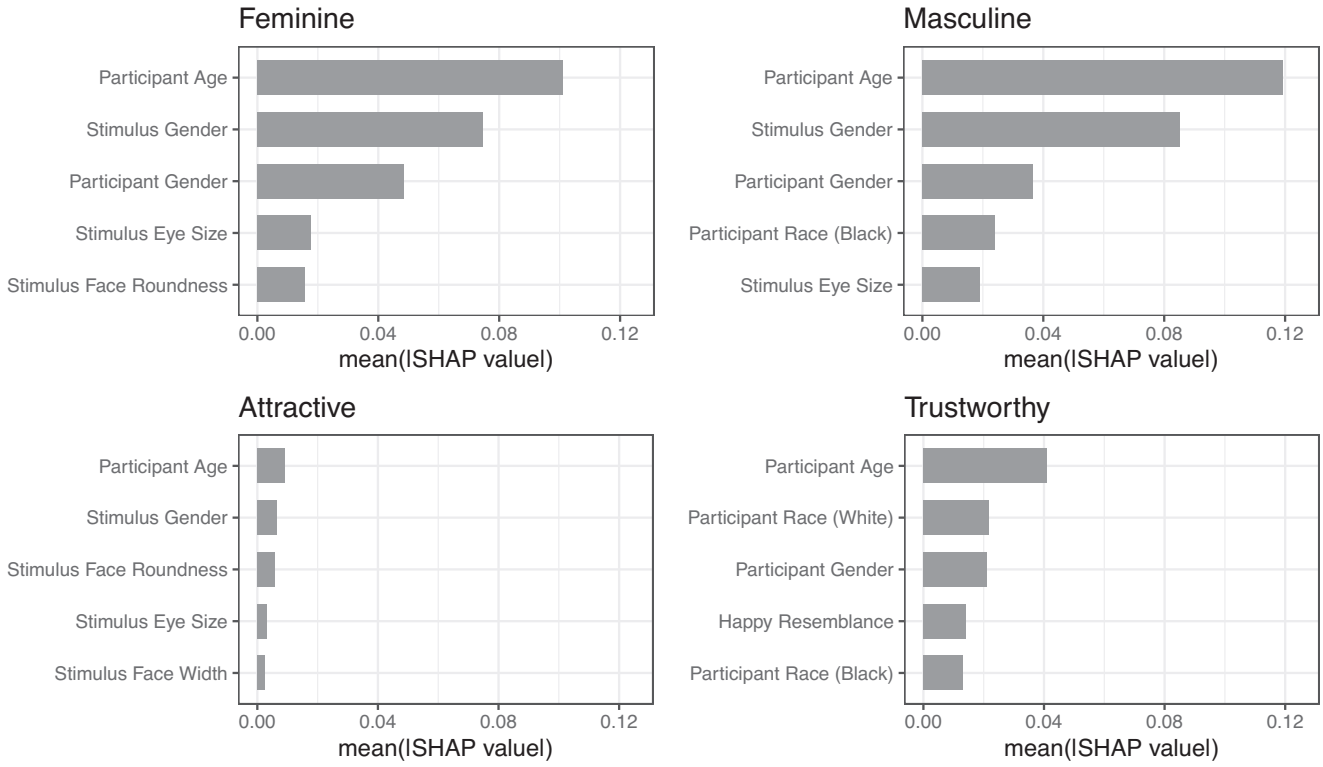
judgments of less diverse stimuli. This effect was independent of the effect of the type of judgment.

The second goal of this research was to move beyond descriptive analyses and toward a deeper understanding of what predicts the specific variance components of facial judgments. To accomplish this, we selected a number of features theoretically related to face judgments and used them to predict the shared and idiosyncratic regression coefficients using machine learning. Our selected predictors included both stimulus-level features (e.g., skin luminance, face shape, incidental emotion resemblance) and participant-level features (self-reported race, gender, and age). Next, we interrogated our models in order to determine which features were the most powerful predictors. Many of the important features in the shared variance models were what the literature suggests should be related to the judgment being examined. We found that perceived face gender was the most important predictor for the shared preference portion of attractiveness, masculinity, and femininity judgments. Likewise, other typically sexual dimorphic physical cues such as face width, face luminance, and eye size followed in importance, but were less important than perceived face gender. This set of predictive features has been shown to be related to judgments of attractiveness, femininity, and masculinity (e.g., Penton-Voak et al., 2004; Said & Todorov, 2011). Likewise, the probability that a face resembled a happy expression (and not a sad expression) was the most important feature for predicting the shared portion of trustworthiness judgments, again aligning with decades of research that suggests faces judged to be more trustworthy appear happier and more feminine (e.g., Oosterhof & Todorov, 2008; Said et al., 2009). It is important to note that other than perceived gender, other social demographics of the stimuli (i.e., perceived race and age) had little influence on the model's predictions. This observation is important to acknowledge as much research attention has been given to how the social demographics of both the perceiver and the target influence judgments. This observation provides additional evidence that there are little cross-race differences in judgments where one might expect a difference or where there has been mixed evidence (e.g., facial attractiveness; Burke et al., 2013; Kleisner et al., 2017; Langlois et al., 2000; Martinez & Paluck, 2020; Zebrowitz et al., 2012).

Critically, much of the past research that has determined important features used in making facial judgments has examined responses at the group level. Comparatively little is known about what factors are important for predicting how judgments are made from one individual to another. Our idiosyncratic models consistently identified participant age as the most important predictor for individual preference coefficients, even when additional stimulus-level features were included as predictors in the model. While the effect of (old) age is a relatively under-studied area within perception

Figure 5

Mean Absolute SHAP Values (x Axis) for the Top Five Most Important Features of Each Idiosyncratic Variance Model



Note. The SHAP value of each feature can be interpreted as the degree to which that feature, on average, affects the final result of the model. Values are additive, meaning that each SHAP value adds that amount to the model's final predicted value relative to the model's average prediction. It should be noted that the SHAP values for the idiosyncratic models are much smaller than those of the shared models. SHAP = Shapley additive explanation.

and judgment compared to race and gender, there is evidence that some judgments are influenced by the perceiver's age, such as judgments of attractiveness (He et al., 2021). Further, older adults are also more likely to report higher levels of trust regardless of who they are evaluating compared to younger adults (Bailey & Leon, 2019). This finding underscores that aside from age, other participant demographics measured provide relatively little to the prediction, at least in the present context. One notable exception to this is that participant gender was identified as an important predictor for idiosyncratic judgments of femininity and masculinity, which is in line with past research that suggests women and men weight femininity and masculinity differently when making certain evaluations, such as attractiveness and trustworthiness (e.g., DeBruine et al., 2010; He et al., 2021; Mattarozzi et al., 2015).

One limitation to acknowledge is that our idiosyncratic models performed significantly worse than our shared models, particularly for judgments of attractiveness and trustworthiness. But this is to be expected, as the idiosyncratic coefficients we are predicting are by definition the portion of the response that remains after accounting for variance attributable to the stimulus and participant. There is a rich potential for future research to examine additional predictors that might be influencing idiosyncratic preferences. For example, other social information such as political orientation, socioeconomic status, anxious or depressive symptomatology, and life satisfaction among many others might add to a more complete understanding of how one individual forms their judgment in relation to others.

It should be noted that our idiosyncratic models are still averaging predictors to some degree (e.g., the model's tree-based approach determined that all participants within an age range will respond a certain way to specific stimuli), much like research that uses individual differences to predict individual responses aggregates predictors to varying degrees. In contrast, the participant-by-stimulus interaction could also be interpreted as an irreducible idiosyncrasy that is specific to the person and stimulus pair that cannot be explained by a more general variable-focused analysis. This is an important distinction to consider as this interpretation would suggest that very little (if any) variance explained by the interaction component could be predicted from features that are not specific and unique to the individual judging a particular stimulus. The best approach for examining idiosyncrasies without any averaging would be to build individual participant-level models of judgments and determine how effective each participant's model is at predicting their responses. This would be an interesting avenue for future work to explore, as current research has yet to tease apart and compare the predictive power of different averaged and idiosyncratic variables within the interaction component itself, as the participant-by-stimulus interaction is likely made up of some weighted combination of both "group averaged idiosyncrasies" as well as irreducible idiosyncrasies. One recent approach has utilized generative modeling to visualize photorealistic, individualized models of judgments of faces (Albohn et al., 2022, 2024; Todorov et al., 2023). As technology advances, the resolution and precision of individualized models will increase in predictive power allowing for a

more nuanced understanding of important aggregate (at any level of aggregation) and irreducible idiosyncratic predictors.

Taken together, the present work provides clear evidence that the amount of variance attributable to stimuli or participants is not a fixed estimate. Rather, attributable variance appears to be influenced by a number of factors, such as the amount of ambiguity present in the judgment being made, the stimuli selected to solicit judgments, or the scale used to measure it. This has both theoretical and practical importance, such as identifying the sources of variance to more closely align with consumer priorities (Govers & Schoormans, 2005; Tarka et al., 2022) and examining heterogeneity—particularly participant idiosyncrasies—in causal processes that have the potential to alter research outcomes (Bolger et al., 2019).

Conclusion

Human judgment is a complex process. While there is certainly some shared agreement in facial judgments between individuals, there also appears to be just as much idiosyncrasy. What one individual finds attractive, trustworthy, feminine, or masculine is not the same as every other individual. Even when people agree, they may utilize different facial features or weigh each feature differently in order to arrive at their judgment. The current research represents an incremental, but important, step forward for identifying what does and does not influence individualized and group-level judgments. Knowing this, researchers can begin to more accurately identify potential variables of interest and build theoretically stronger models that explain how individuals make decisions and what influences their preferences.

References

Adams, R. B., Albohn, D. N., Hedgecoth, N., Garrido, C. O., & Adams, K. D. (2022). Angry White faces: A contradiction of racial stereotypes and emotion-resembling appearance. *Affective Science*, *3*(1), 46–61. https://doi.org/10.1007/s42761-021-00091-5

Albohn, D. N., & Adams, R. B. (2021). The expressive triad: Structure, color, and texture similarity of emotion expressions predict impressions of neutral faces. *Frontiers in Psychology*, *12*, Article 612923. https://doi.org/10.3389/fpsyg.2021.612923

Albohn, D. N., Uddenberg, S., & Todorov, A. (2022). A data-driven, hyper-realistic method for visualizing individual mental representations of faces. *Frontiers in Psychology*, *13*, Article 997498. https://doi.org/10.3389/fpsyg.2022.997498

Albohn, D. N., Uddenberg, S., & Todorov, A. (2024). *Individualized models of social judgments and context-dependent representations*. PsyArXiv. https://doi.org/10.31234/osf.io/ug92m

Albright, L., Kenny, D. A., & MaUoy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Interpersonal Relations and Group Processes*, *55*(3), 387–395. https://doi.org/10.1037//0022-3514.55.3.387

Bailey, P. E., & Leon, T. (2019). A systematic review and meta-analysis of age-related differences in trust. *Psychology and Aging*, *34*(5), 674–685. https://doi.org/10.1037/pag0000368

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bjornsdottir, R. T., Hehman, E., & Human, L. J. (2022). Consensus enables accurate social judgments. *Social Psychological and Personality Science*, *13*(6), 1010–1021. https://doi.org/10.1177/19485506211047095

Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General*, *148*(4), 601–618. https://doi.org/10.1037/xge0000558

Burke, D., Nolan, C., Hayward, W. G., Russell, R., & Sulikowski, D. (2013). Is there an own-race preference in attractiveness? *Evolutionary Psychology*, *11*(4), 855–872. https://doi.org/10.1177/147470491301100410

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

DeBruine, L. M., Jones, B. C., Smith, F. G., & Little, A. C. (2010). Are attractive men's faces masculine or feminine? The importance of controlling confounds in face stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(3), 751–758. https://doi.org/10.1037/a0016457

Diego-Mas, J. A., Fuentes-Hurtado, F., Naranjo, V., & Alcañiz, M. (2020). The influence of each facial feature on how we perceive and interpret human faces. *i-Perception*, *11*(5), Article 204166952096112. https://doi.org/10.1177/2041669520961123

Ebner, N. C. (2008). Age of face matters: Age-group differences in ratings of young and old faces. *Behavior Research Methods*, *40*(1), 130–136. https://doi.org/10.3758/BRM.40.1.130

Germine, L., Russell, R., Bronstad, P. M., Blokland, G. A. M., Smoller, J. W., Kwok, H., Anthony, S. E., Nakayama, K., Rhodes, G., & Wilmer, J. B. (2015). Individual aesthetic preferences for faces are shaped mostly by environments, not genes. *Current Biology*, *25*(20), 2684–2689. https://doi.org/10.1016/j.cub.2015.08.048

Govers, P. C. M., & Schoormans, J. P. L. (2005). Product personality and its influence on consumer preference. *Journal of Consumer Marketing*, *22*(4), 189–197. https://doi.org/10.1108/07363760510605308

Hauser, D. J., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2023). Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. *Behavior Research Methods*, *55*(8), 3953–3964. https://doi.org/10.3758/s13428-022-01999-x

He, D., Workman, C. I., Kenett, Y. N., He, X., & Chatterjee, A. (2021). The effect of aging on facial attractiveness: An empirical and computational investigation. *Acta Psychologica*, *219*, Article 103385. https://doi.org/10.1016/j.actpsy.2021.103385

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., Xie, S. Y., & Hehman, E. (2021). Little between-region and between-country variance when people form impressions of others. *Psychological Science*, *32*(12), 1907–1917. https://doi.org/10.1177/0956797621101995O

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(2), 199–209. https://doi.org/10.1037/0096-1523.32.2.199

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., … Coles, N. A. (2021). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour*, *5*(1), 159–169. https://doi.org/10.1038/s41562-020-01007-2

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. William Collins.

Karkkainen, K., & Joo, J. (2021). *FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation*. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1547–1557). https://doi.org/10.1109/WACV48630.2021.00159

Kenny, D. A., & La Voie, L. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 18, pp. 141–182). Elsevier. https://doi.org/10.1016/S0065-2601(08)60144-6

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758.

Kleisner, K., Kočnar, T., Tureček, P., Stella, D., Akoko, R. M., Třebický, V., & Havlíček, J. (2017). African and European perception of African female attractiveness. *Evolution and Human Behavior*, *38*(6), 744–755. https://doi.org/10.1016/j.evolhumbehav.2017.07.002

Kurosu, A., & Todorov, A. (2017). The shape of novel objects contributes to shared impressions. *Journal of Vision*, *17*(13), Article 14. https://doi.org/10.1167/17.13.14

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, *126*(3), 390–423. https://doi.org/10.1037/0033-2909.126.3.390

Leppänen, J. M., Milders, M., Bell, J. S., Terriere, E., & Hietanen, J. K. (2004). Depression biases the recognition of emotionally neutral faces. *Psychiatry Research*, *128*(2), 123–133. https://doi.org/10.1016/j.psychres.2004.05.020

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4768–4777). Curran Associates.

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

Malloy, C. S., Hughes, C., & Cassidy, B. S. (2023). Perceiver and target partisanship shift facial trustworthiness effects on likability. *Scientific Reports*, *13*(1), Article 6130. https://doi.org/10.1038/s41598-023-33307-8

Martinez, J. E. (2023). Facecraft: Race reification in psychological research with faces. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916231194953

Martinez, J. E., Funk, F., & Todorov, A. (2020). Quantifying idiosyncratic and shared contributions to judgment. *Behavior Research Methods*, *52*(4), 1428–1444. https://doi.org/10.3758/s13428-019-01323-0

Martinez, J. E., & Paluck, E. L. (2020). *Quantifying shared and idiosyncratic judgments of racism in social discourse* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/kfpjg

Mattarozzi, K., Todorov, A., Marzocchi, M., Vicari, A., & Russo, P. M. (2015). Effects of gender and personality on first impression. *PLoS ONE*, *10*(9), Article e0135529. https://doi.org/10.1371/journal.pone.0135529

Oh, D., Dotsch, R., Porter, J., & Todorov, A. (2020). Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology: General*, *149*(2), 323–342. https://doi.org/10.1037/xge0000638

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Paunonen, S. V., Ewan, K., Earthy, J., Lefave, S., & Goldberg, H. (1999). Facial features as personality cues. *Journal of Personality*, *67*(3), 555–583. https://doi.org/10.1111/1467-6494.00065

Penton-Voak, I. S., Jacobson, A., & Trivers, R. (2004). Populational differences in attractiveness judgements of male and female faces. *Evolution and Human Behavior*, *25*(6), 355–370. https://doi.org/10.1016/j.evolhumbehav.2004.06.002

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, *119*(17), Article e2115228119. https://doi.org/10.1073/pnas.2115228119

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

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, *57*(1), 199–226. https://doi.org/10.1146/annurev.psych.57.102904.190208

Rhodes, G., Proffitt, F., Grady, J. M., & Sumich, A. (1998). Facial symmetry and the perception of beauty. *Psychonomic Bulletin & Review*, *5*(4), 659–669. https://doi.org/10.3758/BF03208842

Roth, W. D. (2016). The multiple dimensions of race. *Ethnic and Racial Studies*, *39*(8), 1310–1338. https://doi.org/10.1080/01419870.2016.1140793

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, *9*(2), 260–264. https://doi.org/10.1037/a0014681

Said, C. P., & Todorov, A. (2011). A statistical model of facial attractiveness. *Psychological Science*, *22*(9), 1183–1190. https://doi.org/10.1177/0956797611419169

Serengil, S. I., & Ozpinar, A. (2021). *HyperExtended LightFace: A facial attribute analysis framework*. 2021 International Conference on Engineering and Emerging Technologies (ICEET) (pp. 1–4). https://doi.org/10.1109/ICEET53442.2021.9659697

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*(3), 349–354. https://doi.org/10.1177/0956797610362647

Sutherland, C. A. M., Burton, N. S., Wilmer, J. B., Blokland, G. A. M., Germine, L., Palermo, R., Collova, J. R., & Rhodes, G. (2020). Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proceedings of the National Academy of Sciences*, *117*(19), 10218–10224. https://doi.org/10.1073/pnas.1920131117

Tarka, P., Kukar-Kinney, M., & Harnish, R. J. (2022). Consumers' personality and compulsive buying behavior: The role of hedonistic shopping experiences and gender in mediating-moderating relationships. *Journal of Retailing and Consumer Services*, *64*, Article 102802. https://doi.org/10.1016/j.jretconser.2021.102802

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press.

Todorov, A., & Oh, D. (2021). The structure and perceptual basis of social judgments from faces. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 63, pp. 189–245). Elsevier. https://doi.org/10.1016/bs.aesp.2020.11.004

Todorov, A., Suchow, J. W., Peterson, J., Uddenberg, S., & Griffiths, T. L. (2022). *Reflections on ML models of first impressions*. https://medium.com/@pnas.2115228119/reflections-on-ml-models-of-first-impressions-5bd28d222ff6

Todorov, A., Uddenberg, S., & Albohn, D. N. (2023). Generative models for visualizing idiosyncratic impressions. *British Journal of Psychology*, *114*(2), 511–514. https://doi.org/10.1111/bjop.12622

Welling, L. L. M., DeBruine, L. M., Little, A. C., & Jones, B. C. (2009). Extraversion predicts individual differences in women's face preferences. *Personality and Individual Differences*, *47*(8), 996–998. https://doi.org/10.1016/j.paid.2009.06.030

Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science*, *26*(3), 237–242. https://doi.org/10.1177/0963721416683996

Zebrowitz, L. A., Wang, R., Bronstad, P. M., Eisenberg, D., Undurraga, E., Reyes-García, V., & Godoy, R. (2012). First impressions from faces among U.S. and culturally isolated Tsimane' people in the Bolivian rainforest. *Journal of Cross-Cultural Psychology*, *43*(1), 119–134. https://doi.org/10.1177/0022022111411386

Zhan, J., Liu, M., Garrod, O. G. B., Daube, C., Ince, R. A. A., Jack, R. E., & Schyns, P. G. (2021). Modeling individual preferences reveals that face beauty is not universally perceived across cultures. *Current Biology*, *31*(10), 2243–2252.e6. https://doi.org/10.1016/j.cub.2021.03.013

Zietsch, B. P., Lee, A. J., Sherlock, J. M., & Jern, P. (2015). Variation in women's preferences regarding male facial masculinity is better explained by genetic differences than by previously identified context-dependent effects. *Psychological Science*, *26*(9), 1440–1448. https://doi.org/10.1177/0956797615591770

Received January 1, 2024
Revision received May 20, 2024
Accepted June 17, 2024 ■