American Psychological Association

Journal of Experimental Psychology: General

© 2019 American Psychological Association
0096-3445/20/$12.00

2020, Vol. 149, No. 2, 323–342
http://dx.doi.org/10.1037/xge0000638

Gender Biases in Impressions From Faces: Empirical Studies and Computational Models

DongWon Oh
Princeton University

Ron Dotsch
Utrecht University

Jenny Porter and Alexander Todorov
Princeton University

Trustworthiness and dominance impressions summarize trait judgments from faces. Judgments on these key traits are negatively correlated to each other in impressions of female faces, implying less differentiated impressions of female faces. Here we test whether this is true across many trait judgments and whether less differentiated impressions of female faces originate in different facial information used for male and female impressions or different evaluation of the same information. Using multidimensional rating datasets and data-driven modeling, we show that (a) impressions of women are less differentiated and more valence-laden than impressions of men and find that (b) these impressions are based on similar visual information across face genders. Female face impressions were more highly intercorrelated and were better explained by valence (Study 1). These intercorrelations were higher when raters more strongly endorsed gender stereotypes. Despite the gender difference, male and female impression models—derived from separate trustworthiness and dominance ratings of male and female faces—were similar to each other (Study 2). Further, both male and female models could manipulate impressions of faces of both genders (Study 3). The results highlight the high-level, evaluative effect of face gender in impression formation—women are judged negatively to the extent their looks do not conform to expectations, not because people use different facial information across genders but because people evaluate the information differently across genders.

Keywords: face perception, social perception, social cognition, gender stereotypes

Supplemental materials: http://dx.doi.org/10.1037/xge0000638.supp

People effortlessly attribute traits, such as competence and emotional stability, to others based on their facial appearance (Todorov, 2017; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). These impressions of traits affect a variety of important real-world outcomes, which range from voting behavior (Antonakis & Dalgas, 2009; Lenz & Lawson, 2011; Little, Burriss, Jones, & Roberts, 2007; Olivola & Todorov, 2010; Todorov, Mandisodza, Goren, & Hall, 2005), to court decisions (Blair, Judd, & Chapleau, 2004; Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006; Wilson & Rule, 2015; Zebrowitz & McDonald, 1991), to mating choices (Cooper, Dunne, Furey, & O'Doherty, 2012). Because the trait impressions are highly intercorrelated, by examining the relations among the perceived traits, researchers can succinctly describe the structure of face impression formation (Oosterhof & Todorov, 2008; Sutherland et al., 2013), and then using the impression structure, predict impressions on multiple traits and reveal the facial information underlying these impressions (Oh, Buck, & Todorov, 2019; Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013; Walker & Vetter, 2009, 2016).

There is a large body of research on the structure underlying the relations between traits in impressions (Fiske, Cuddy, Glick, & Xu, 2002; Imhoff & Koch, 2017; Koch, Imhoff, Dotsch, Unkelbach, & Alves, 2016; Oosterhof & Todorov, 2008; Osgood, Suci, & Tannenbaum, 1957; Rosenberg, Nelson, & Vivekananthan, 1968; Wiggins, 1979). In face-based first impressions, the impressions are reducible to a small number of summary dimensions—valence and physical power—approximated by judgments of trustworthiness and dominance (Oosterhof & Todorov, 2008; Walker & Vetter, 2016; but see Sutherland et al., 2013, which found three dimensions, including the new dimension of attractiveness). Importantly, the structure of impressions may vary across meaningful subcategories of faces, such as men and women. However, previous research on face impressions has mostly ne-

This article was published Online First July 11, 2019.

DongWon Oh, Department of Psychology, Princeton University; Ron Dotsch, Department of Psychology, Utrecht University; Jenny Porter and Alexander Todorov, Department of Psychology, Princeton University.

Ron Dotsch is now at Anchorman, Amsterdam, the Netherlands. Jenny Porter is now at Clayton Early Learning, Denver, CO.

All stimuli and datasets are available at Open Science Framework: https://osf.io/ycv72/.

Correspondence concerning this article should be addressed to DongWon Oh, who is now at the Department of Psychology, New York University, 6 Washington Place, Room 454, New York, NY 10006. E-mail: dong.w.oh@gmail.com

glected potential gender differences, implicitly assuming the same structure of impressions across genders. This assumption is inconsistent with both empirical evidence and theoretical reasoning.

Gender Biases in Impressions

Empirical data suggest that impressions from facial appearance are more highly correlated for women than for men. Trustworthiness and dominance impressions, for instance, are negatively correlated for female faces, but not for male faces (Sutherland, Young, Mootz, & Oldmeadow, 2015). Dominant female faces are perceived more negatively than nondominant female faces, nondominant male faces, and dominant male faces. These findings are inconsistent with the existing model of face trait attribution, which assumes that the two summary dimensions of valence/trustworthiness and power/dominance are orthogonal to each other (see Rosenberg et al., 1968, for the model of nonvisual person perception, where these dimensions are correlated).[1] Given the high correlations of these two trait impressions with other trait impressions (Oosterhof & Todorov, 2008; Sutherland et al., 2013), it is likely that face impressions overall are more highly intercorrelated for women than for men.[1]

The idea of less differentiated face impressions of women aligns well with the rich literature on gender stereotypes. People expect men and women to be and behave in certain ways (e.g., women to be more submissive, dependent, and gentle than men; Bem, 1974; I. K. Broverman, Vogel, Broverman, Clarkson, & Rosenkrantz, 1972; Prentice & Carranza, 2002; Spence, Helmreich, & Holahan, 1979; Spence, Helmreich, & Stapp, 1975). These beliefs are widely held across cultures (Williams & Best, 1990) and difficult to change (Prentice & Carranza, 2004). Although the gender-associated expectations are held for both men and women, there are a larger number of traits that are considered typical or desirable for women than for men (Prentice & Carranza, 2002), including valence-related traits such as kindness and friendliness (Heilman, 2001; Rudman, 1998; Rudman & Glick, 2001). Relatedly, women are evaluated positively to the extent that they conform to the stereotypes associated with them (benevolent sexism; Glick & Fiske, 1996; Glick et al., 2000) unlike men, who are freer from the normative boundaries of stereotypes. If this principle applies to facial impressions, women whose appearances suggest traits inconsistent with the stereotypes (e.g., a woman with a face that makes other people intuitively judge that she has a domineering personality) are likely to be evaluated more negatively than men with the same degree of stereotype-inconsistency in their appearance. It would follow that in females' impressions, more traits should be correlated with valence than in males' impressions, consistent with the previously found negative correlation between perceived facial dominance and trustworthiness for women (Sutherland et al., 2015). For these reasons, we expect less differentiated face impressions for female than for male faces.

Notably, the social cognition theories mentioned above (e.g., benevolent sexism) are about what behavior and traits people expect of others and do not make a direct prediction about how people evaluate faces. However, facial appearance serves as a source of trait inferences that in people's minds are predictive of behavior: People effortlessly judge a person's traits from their face with a high level of within- and cross-rater reliability and act on these inferences (for review, see Todorov et al., 2015). Moreover, some facial appearances reliably lead to trait inferences that are stereotypical (or counterstereotypical) given the person's social category. In this article, we refer to such an appearance as a "stereotypical (or counterstereotypical) appearance." For example, a woman with masculine facial features (e.g., strong chin) is inferred to be dominant, and thus could be described as having a "counterstereotypical appearance" in this sense. These kinds of inferences are influenced by both low-level visual information (e.g., masculine facial features) and high-level category inferences (e.g., gender expectations such as women are not dominant).

The Present Research

Using dimensionality reduction and computational modeling, the current article examines the principles behind the gender difference in facial impressions, expanding prior research in two ways. First, by examining the degree of intercorrelations between the ratings of face impressions, we investigate whether and to what extent female face impressions are less differentiated than male face impressions. Across three datasets, we find that female impressions are indeed less differentiated. Further, by measuring a rater characteristic related to social perception and expectations, namely, the degree to which the raters endorse gender stereotypes, we test whether the degree of impression differentiation is related to how much raters endorse gender stereotypes. Specifically, if the gender-related difference in impression differentiation stems from gender stereotypes as we argue, then perceivers who endorse gender stereotypes more strongly would show less differentiated impressions of faces.

Second, by building separate data-driven computational models of impressions (Oosterhof & Todorov, 2008) of male and female faces and cross-validating the models across face genders, we investigate what is at the basis of the less differentiated impressions for women. Specifically, the gender difference in impression differentiation can stem from either (a) the same visual information used differently across genders to form an impression or (b) different visual information used to form the impression. For facial attractiveness, for example, the same visual information is used across genders but has the opposite evaluative outcome, in which masculine face reflectance increases the attractiveness of men but decreases the attractiveness of women (Said & Todorov, 2011). This supports the first hypothesis. However, it is yet to be tested whether this would generalize to key face impressions such as trustworthiness and dominance.

A data-driven face model of a trait impression (e.g., competence) represents what visual information people use to form the impression (Dotsch & Todorov, 2012; Jack & Schyns, 2017; Todorov, Dotsch, Wigboldus, & Said, 2011). If people use the same information when forming impressions of male and female faces (e.g., masculine facial properties to infer dominance), then the models of male and female impressions should be similar. Such a result would imply that differences between male and female

[1] Trustworthiness, in social-perception and gender-study literatures, is also referred to as *communion* (Abele, 2003; Cislak & Wojciszke, 2008), *warmth* (Fiske et al., 2002), or *approachability* (Sutherland et al., 2013). These concepts are highly similar in people's minds (e.g., Sutherland, Oldmeadow, & Young, 2016). To avoid confusion, the current article uses the word *valence* to represent the positivity/negativity in social evaluation.

impression differentiation are caused by a different evaluative process resulting from gender categorization. Categorizing a face as female, for example, would lead to stronger correlations between impressions that reflect female gender stereotypes. On the other hand, if people use different information when forming impressions of male and female faces, then models of male and female impressions should be different.

We can test these possibilities by (a) looking directly at the similarity of the male and female models and (b) cross-validating the models of impressions on novel male and female faces. To the extent that the models for male and female faces are similar, the models would be highly correlated, and their effects on impressions would be similar irrespective of whether they are applied to male or female faces (e.g., impressions of the trustworthiness of a novel female face would be similar whether the face is manipulated by a male or a female model of trustworthiness). In contrast, even if the models for male and female faces are highly correlated, to the extent that they are based on different information, they would have different effects on impressions depending on whether they are applied to a male or a female face (e.g., impressions of the trustworthiness of a novel female face would be more successfully manipulated by a female than by a male model of trustworthiness).

Before reporting individual studies, we would like to make a clear distinction between (a) the outcome of social perception, and the potential (b) low- and (c) high-level mechanisms underlying this outcome. In the case of the gender difference in facial impression differentiation, Study 1 investigates whether female impressions are less differentiated and more valence-laden than male impressions, findings that would imply counterstereotypical looks in women are negatively evaluated (the outcome). Facial appearance per se admittedly cannot be stereotypical or counterstereotypical, because a stereotype is described and prescribed at the level of traits and behaviors. However, as explained above, facial appearance in people's minds serves as a reliable source of trait inferences (e.g., assertiveness, tenderness) and expectations of behaviors (e.g., loud voice, gentle bodily gestures). Studies 2 and 3 investigate the potential mechanisms underlying the findings (less differentiated impressions in women's faces): Do people use different visual information when forming male and female facial impressions (a low-level mechanism) or do people interpret visual information differently across genders (a high-level mechanism)? Our findings suggest it is the latter.

Study 1: Impression Differentiation in Male and Female Faces

In Study 1, we compare two measures between genders to assess potential differences in the level of impression differentiation: (a) the degree to which multiple impressions in each gender are intercorrelated and (b) the degree to which specific impressions are explained by general valence of impressions. We expected less differentiated face impressions for women than for men, expressed in (a) a stronger correlation between specific trait impressions for female than for male faces, and (b) a larger variance explained by the first principal component (PC1) for female than for male faces to the extent that PC1 captures the valence of impressions.

Study 1a: Reanalysis of Oosterhof and Todorov (2008)

In Study 1a, we analyze preexisting rating datasets of male and female face images. In the original work, Oosterhof and Todorov (2008) conducted a PCA on the rating dataset without consideration of face gender and found that the first two principal components, which could be interpreted as valence and power, accounted for over 80% of the variance.

Method.

Participants. Three hundred and one Princeton University undergraduate students were recruited by Oosterhof and Todorov (2008) and participated in the trait rating experiments for partial course credit or cash.

Stimuli. Sixty-six (33 males, 33 females) naturalistic face photos with direct gaze and resting expressions were used (Lundqvist, Flykt, & Arne, 1998). The individuals in the photos were White amateur actors between the ages of 20–30 with no facial hair, earrings, eyeglasses, or visible make-up, all wearing gray T-shirts.

Procedure. Participants rated 33 male and 33 female face photos on 14 traits—how aggressive, attractive, caring, confident, dominant, emotionally stable, intelligent, mean, responsible, sociable, threatening, trustworthy, unhappy, or weird each individual looked. These traits were selected due to their empirical and theoretical importance: The traits (except for dominance) explained about 68% of unconstrained, spontaneous person descriptions from face images (Oosterhof & Todorov, 2008). Dominance was included because of its importance in personality perception (Wiggins, 1979).

To collect the face impression ratings, different groups of participants were assigned to form impressions of all 66 faces on a single trait ($n_{\text{rater}} \geq 18$). That is, each participant rated the faces on a single trait. Participants were told that the study was about first impressions and were encouraged to rely on their "gut feeling." The faces were presented and rated in three separate blocks to reduce the measurement error for each participant's answers. The average face rating of each participant served as the measure of their evaluation on the respective trait. This procedure also allows for screening out unreliable raters—those who show zero or negative test-retest within-rater reliability as calculated between ratings in different blocks.

Each face image was presented in color at the center of the screen (220 × 298 pixels with the height of the face being about 206 pixels) with a question above the face ("How [*trait term*] is this person?") and a response scale below the face ("1 Not at all [*trait term*] - 9 Extremely [*trait term*]"). Each face was visible until the participant responded, the intertrial interval (ITI) was 1,000 ms, and the order of faces was randomized. All 14 trait ratings showed moderate to high interrater agreement ($r_{\min} = .26$) and interrater reliability ($\alpha_{\min} = .90$). To obtain impression measures for each face, the ratings on the 14 traits were averaged across raters. The mean rating dataset is available at Open Science Framework: https://osf.io/ycv72/.

We conducted two main analyses to test for gender differences in the level of differentiations in impressions. In the first analysis, we calculated the extent to which traits are correlated to one another in each gender and compared the level of correlations between ratings of male and female faces. Specifically, we computed pairwise correlational coefficients among all 14 trait ratings

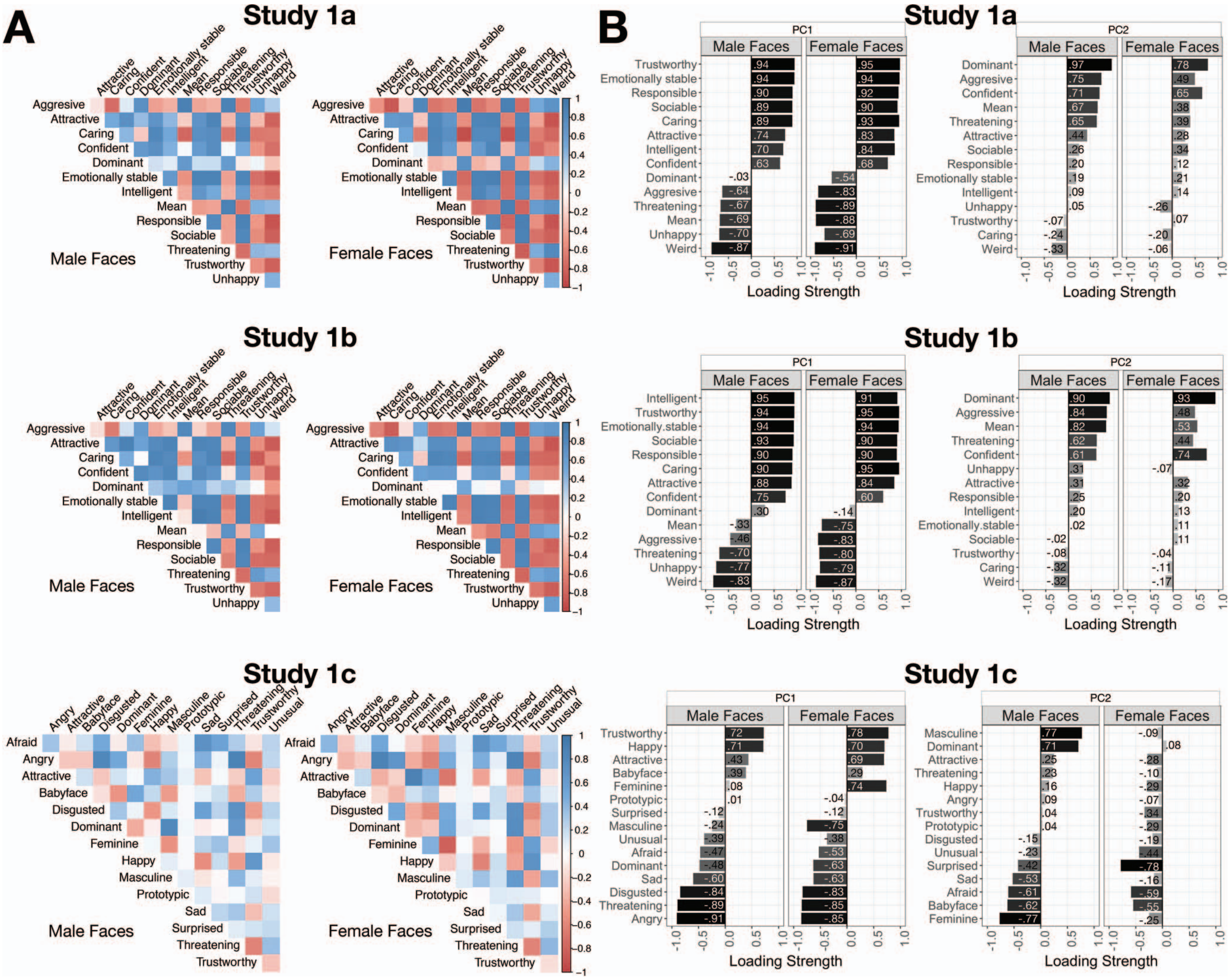

Figure 1. Correlational analyses and principal component analysis (PCA) results for male and female faces in Study 1. Female impressions are less differentiated than male impressions (A), and female impressions load more highly on the first principal component (PC1; valence) than male impressions do (B). For the correlational analysis, a Pearson correlational matrix was computed between all trait pairs within each face gender. The color of each cell represents the strength of the correlation (darker: stronger correlation; A). For PCA, an orthogonal PCA was conducted for each face gender. The magnitude of each bar and the number on each bar represents the loading of the respective trait on PC1 (left) and second principal component (PC2) (right) for each face gender. The traits are sorted in the order of the loading strength in male faces ratings (B).

($_{14}C_2$ = 91 pairs) separately for male and female faces (Figure 1A, top). Because we were interested in contrasting the average strength of interimpression correlation between face genders, we converted all coefficients into positive values. To test whether female facial impressions were more strongly correlated, or less differentiated, to each other than male facial impressions were, we conducted a test of the matrix equality between the two gender-specific matrices (Jennrich, 1970). Higher absolute values of the correlational coefficients for female than male faces, along with a significant difference in the correlation matrices, would implicate higher dependency between perceptions of traits for female faces.

In the second analysis, we *z*-transformed the average trait ratings within each trait in each face gender, and then subjected the ratings to an orthogonal PCA for each gender. No rotation was made. We reported and visualized the components with eigenvalues bigger than 1 (the Kaiser rule), following the original study by Oosterhof and Todorov (Supplemental Table S1 in the online supplemental material). To test whether female facial impressions were more driven by valence than male facial impressions, we compared the amount of the variance explained by PC1 for each gender. A larger variance explained by PC1 would implicate greater dependency of impressions on valence to the extent that PC1 is loaded highly on by valenced impression ratings (e.g., responsible, mean).

Results and discussion. Consistent with the PCA results collapsing across male and female faces (Oosterhof & Todorov, 2008), the gender-specific PCAs revealed two key components (see Supplemental Table S1 in the online supplemental material for the eigenvalues and the variance explained by each component). For both genders, PC1 was highly loaded on by all positive (e.g., trustworthy, responsible) or negative traits (e.g., threatening, weird; Figure 1B, top). This is consistent with previous models of face impressions, in which the first component is summarized as valence (Oosterhof & Todorov, 2008) or approachability (Sutherland et al., 2015).

New to our data, female face impressions were less differentiated and more valence-laden than male face impressions (Figures 2A and 2B, top): When the cross-impression correlational matrices were compared between face genders using their absolute values (Jennrich test of matrix equality), the correlation was significantly different, with female ratings being more strongly intercorrelated ($M_{|r|} = 0.68$, $SD_{|r|} = 0.19$) than male ratings ($M_{|r|} = 0.55$, $SD_{|r|} = 0.24$), $\chi^2(91) = 384.37$, $p < .001$ (Figure 2A, top), indicating a higher level of dependency between impressions of female faces. This is consistent with a visual inspection of the PCA solutions: The trait loadings on PC1 have bigger absolute values for female than for male faces and the loadings on PC2 have smaller absolute values for female than for male faces, as shown by longer and darker bars for female faces on average (Figures 1B, top). Correspondingly, the amount of variance explained by PC1, a proxy for valence, was larger for female than male ratings (71.69% vs. 58.40%; Figure 2B, top; Supplemental Table S1 in the online supplemental material), indicating a higher level of dependency of impressions on overall positivity/negativity in female face impressions.

Study 1b: Analysis of a New Dataset

Study 1a revealed that impressions of women are less differentiated and more valence-laden than impressions of men. The objective of Study 1b was to replicate these findings and to test whether the differentiation in impressions is related to relevant stereotypes held by perceivers. Specifically, we expected that the more strongly a perceiver endorses conventional beliefs about genders, the more likely their ratings will show (a) less differentiated and (b) more highly valence-laden impressions. However, whether the effect of stereotype endorsement is stronger for impressions of women's than men's faces or independent of the effect of the face gender on impression differentiation is unclear. To test these hypotheses, we collected new impression ratings of male and female faces and measured participants' level of gender stereotype endorsement.

We also tested whether the gender of participants is related to the gender difference in facial impression differentiation. Because the rater gender was missing in the dataset used in Study 1a, we could not include it in the analyses of Study 1a. In Study 1b, we recorded participants' gender. A possibility is that male raters will show more simplified, valence-laden impressions of women than of men, given prior studies showing stronger endorsement of gender stereotypes by male than female raters (Glick & Fiske, 1996; Swim, Aikin, Hall, & Hunter, 1995; Williams & Best, 1990). Another possibility is that male and female raters will form similar impressions across face genders, given prior studies showing no effect of participant gender on gender-stereotyping (Costrich, Feinstein, Kidder, Marecek, & Pascale, 1975; Deaux & Lewis, 1984; Eagly & Steffen, 1984; Goldberg, 1968; Hagen & Kahn, 1975; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012). A final possibility is that female raters will show more simplified, valence-laden impressions of women than of men, given prior studies showing a higher level of gender-stereotyping (e.g., negative evaluation of women with counterstereotypical traits) by female than male raters (Garcia-Retamero & López-Zafra, 2006; Goldberg, 1968; Parks-Stamm, Heilman, & Hearns, 2008; Rudman, 1998).

Method.

Participants. Five hundred thirty-six online workers living in the United States (258 males, 278 females, one other gender) participated through Amazon Mechanical Turk (MTurk) for monetary reward. Required participant number for each trait was estimated from the interrater reliabilities from Study 1a so that Cronbach's alpha of the ratings would reach .90 for both male and female faces.

Stimuli. Sixty-six (33 males, 33 females) face photos of White individuals used in Study 1a were employed again (Lundqvist et al., 1998).

Procedure. The 14 traits rated in Study 1a were rated by a new group of participants. As in Study 1a, different groups of participants were assigned to form impressions of all 66 faces on a single trait ($n_{\text{rater}} \geq 11$). Participants were given the same instructions as in Study 1a. We asked each participant to judge the faces on a single trait to make the rating task design identical with Study 1a's, avoid participant fatigue, and reduce a possible inflation of correlations between trait ratings (in contrast to a procedure where participant rate the same faces on multiple traits). We kept the same participants from participating in more than one task (e.g., participating in both the "aggressive" rating and the "dominant" rating tasks) using MTurk features. The face stimuli were presented in color at the center of the screen (369 × 500 pixels with the height of the face being about 345 pixels) and were rated twice in separate blocks to reduce the measurement error for each participant's answers: Each participant's ratings were averaged across blocks. As in Study 1a, we calculated each rater's intrarater reliability by correlating their ratings from different blocks. The ratings from participants with zero or negative reliability were excluded, which left us with 469 participants' responses (235 males, 233 females, one other gender).

Each face image was presented at the center of the screen with a question ("How *[trait term]* is this person?") and a response scale below the face, ranging from 1 (Not at all *[trait term]*) to 9 (Extremely *[trait term]*). Each face was visible until the participant responded, the ITI was 1,000 ms, and the order of faces was randomized. All 14 trait ratings showed moderate to high interrater agreement ($r_{\min} = .32$) and interrater reliability ($\alpha_{\min} = .81$). The mean rating dataset is available at Open Science Framework: https://osf.io/ycv72/.

The same two analyses as in Study 1a were conducted (i.e., the correlational analyses and the PCAs) on the ratings averaged across raters to test for gender differences in the differentiation of impressions. First, we compared the degree of the intercorrelations across impressions between genders. Correlational coefficients between the ratings of every trait pair among all 14 traits (${}_{14}C_2 = 91$ pairs) were calculated for each gender. Second, we compared the amount of variance explained by PC1, a proxy of valence.

To test for the effects of the raters' gender and their gender stereotype endorsement, at the end of the study participants were asked to report their gender and to complete a questionnaire regarding gender stereotype endorsement (GSE) that measured the

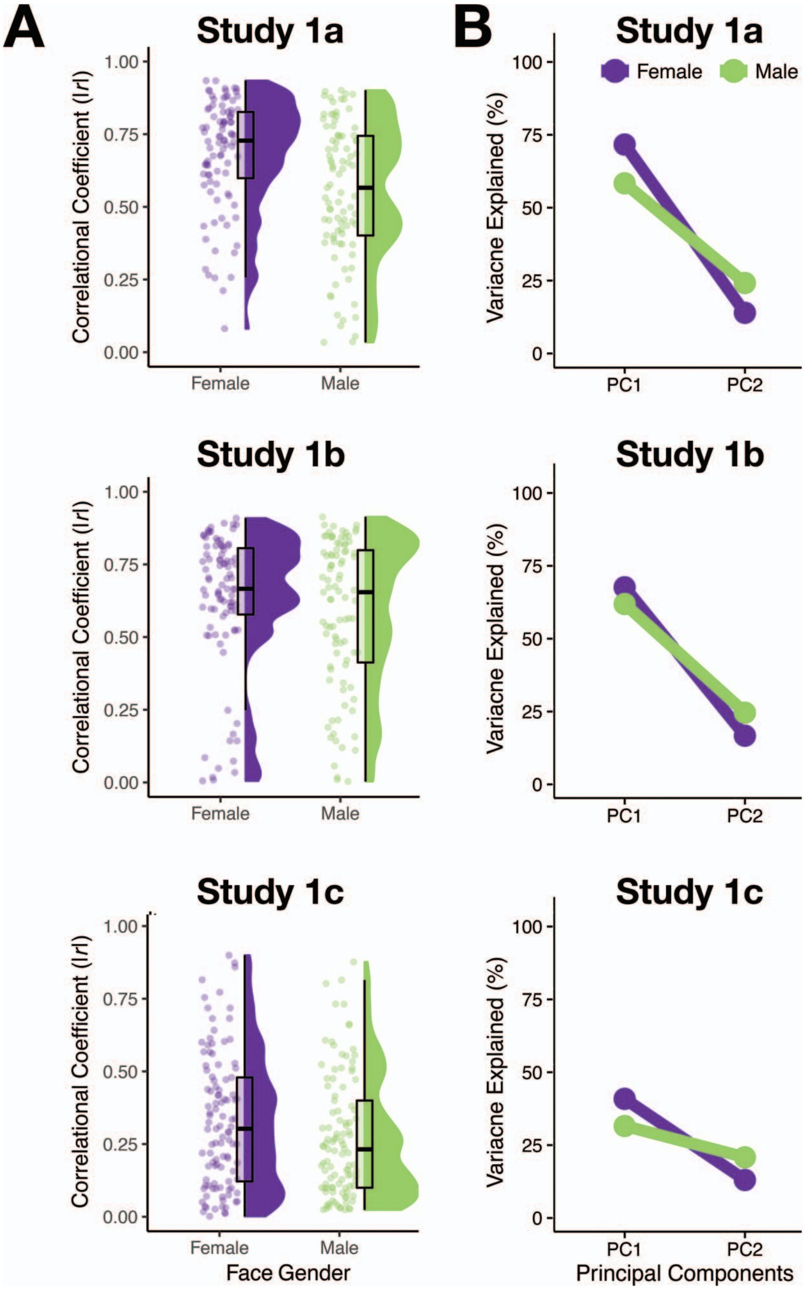

Figure 2. The level of the intercorrelations of trait impressions of men and women (A) and the amount of the variance explained by the first principal component (PC1) and the second principal component (PC2) in the impressions (B). In each study, face-level correlational analyses were conducted between impression ratings separately for male and female face images, and absolute values of the coefficients were compared across genders. Each dot corresponds to the absolute value of the coefficient of the correlation between an impression pair (e.g., the "threatening" and the "unhappy" ratings). The violin plots show the distribution of the values in each face gender, and the dots on the side the raw values. The lower and upper hinges of each box correspond to the 25th and 75th percentiles. The black bar in each box denotes the median. A higher *Y* value represents a lower level of differentiation (or a higher level of intercorrelation) between trait impressions. Across studies, the impressions of women are significantly more highly intercorrelated than the impressions of men (*p*s $< .001$; A). In each study, a PCA was conducted separately for the impressions of male and female images. A higher *Y* value on PC1 represents a stronger relationship between valence of impressions and specific impressions. Across studies, the impressions of women are more valence-laden than the impressions of men (B). PCA = principal component analysis. See the online article for the color version of this figure.

extent to which they agreed with conventional gender stereotypes (Cundiff & Vescio, 2016; Eagly & Mladinic, 1989). Each question ("How do the average man and the average woman compare with each other on how [*trait term*] they are?") was presented with a 9-point response scale ranging from 1 (*Men extremely more*) to 9 (*Women extremely more*). The trait terms were 20 words describing traits either considered stereotypically male and positive (e.g., competitive), or male and negative (e.g., egotistical), female and positive (e.g., nurturing), female and negative (e.g., whiny; Supplemental Table S2 in the online supplemental material for the list). The valence and the gender stereotypicality of these words have been validated (Eagly & Mladinic, 1989; Spence et al., 1979) and the GSE level measured using these words was shown to predict relevant individual characteristics, such as political orientation (Cundiff & Vescio, 2016; Eagly & Mladinic, 1989).

Responses to every question in the questionnaire were significantly different from the middle score (5 in the range of [1,9]) in the stereotype-consistent direction (ts > 7.02, ps $< .001$; Supplemental Figure S1 in the online supplemental material). Half the responses were reverse-coded ($n = 10$) so a larger value meant a higher level of GSE. We replaced each missing value with the weighted response averaged across 10 raters whose other responses were the most similar to the rater's (the *k*-nearest neighbor imputation). Interrater similarity in the responses was determined by the euclidean distance. The GSE questionnaire showed high internal consistency across questions ($\alpha = .90$). Similarly, every item showed a moderately high item-whole correlation (e.g., the correlation between each trait question and the whole questionnaire; $M_r = 0.59$). Based on a high internal consistency across items, we used the sum of the responses to all items as the index of each rater's GSE ("the GSE score"). The GSE score had a possible range of [20,180] with a higher score indicating higher level of stereotype endorsement. To test whether the raters who more strongly endorsed gender stereotypes showed less differentiated facial impressions, we used the rater GSE score to predict the two indices of impression differentiation from the main analyses per face gender: (a) the mean absolute value of interimpression correlational coefficients and (b) the variance explained by PC1. Because we asked each participant to rate every face image only on a single trait (as opposed to all traits), we could not calculate the two indices for each participant. Instead, we subgrouped participants based on an overlapping participant window on the GSE score (see Results for details). We then predicted the two indices using the subgroup's GSE score as a predictor.

Results and discussion. Before reporting the main results, we report how the present data replicate the findings of Oosterhof and Todorov (2008). When the impression ratings were collapsed across face genders, PC1 and PC2 accounted for over 80% of the variance in the ratings newly collected for Study 1b (65.57% and 19.90%, respectively). PC1 and PC2 were loaded highly on by trustworthiness (.95) and dominance ratings (.94), respectively. This replicates Oosterhof and Todorov (2008), where PC1 and PC2 explained over 80% of the variance in the ratings (63.3% and 18.3%, respectively) and were loaded highly on by trustworthiness (.92) and dominance ratings (.87). When the impression ratings were separated by the face gender, again, the PCA solution of each gender replicated Oosterhof and Todorov (2008). Specifically, between the 2008 data and the current data, the component loadings of the impressions were highly similar for both male ($R = .97$) and female face impressions ($R = .98$, see the online supplemental material). In sum, these findings demonstrate high stability of face impressions as the ratings in Oosterhof and Todorov (2008) were collected over 10 years ago.

As in Study 1a, for both genders PC1 was highly loaded on by all positive or negative traits (Figure 1B, middle). This replicates previous models of face impressions (Oosterhof & Todorov, 2008; Sutherland et al., 2013) and Study 1a. Again, female face impressions were less differentiated and more valence-laden than male face impressions. First, the correlational coefficients were significantly different, with female ratings being more strongly intercorrelated ($M_{|r|} = 0.63$, $SD_{|r|} = 0.24$) than male ratings ($M_{|r|} = 0.58$, $SD_{|r|} = 0.26$), $\chi^2(91) = 3145.10$, $p < .001$ (Figure 2A, middle), indicating a higher level of dependency between impressions in female face impressions. Second, PC1 explained a larger amount of variance in female face ratings than in male face ratings (67.94% vs. 61.66%; Figure 2B, middle; Supplemental Table S1 in the online supplemental material). These findings replicate the findings of Study 1a.

Role of raters' gender stereotypes. To examine how the rater's gender stereotypes were related to impression differentiation, we repeated the two analyses (i.e., the correlational analyses and PCAs) for each face gender using responses of multiple rater subsets. We varied the rater subset according to their GSE score. Specifically, starting from the raters who endorsed gender stereotypes the least (i.e., raters with the lowest GSE score, [75,134]) through the raters who endorsed gender stereotypes the most (i.e., raters with the highest GSE score [119,168]), we subgrouped raters based on their GSE score. We performed the analyses using the data from the raters in a window of GSE score with a fixed window width of 49. We then slid the window by 1 GSE score (i.e., increased the start and end points of the window by 1), repeating the correlational analysis and PCA on the ratings averaged across raters per trait in each window. This sliding window's width, the starting point, and the end point were determined so that each window had ≥10 raters per impression trait ($min_n = 245$, $max_n = 410$ in total across traits per window).

We used overlapping windows of raters (rather than nonoverlapping participant subgroups with various levels of GSE) to generate a dependent variable that is continuous. The same analyses with two participant groups divided by the median GSE score (i.e., low- and high-GSE groups) yielded identical results (Supplemental Figure S2 in the online supplemental material).

To statistically compare impression differentiation indices across face genders, compute confidence intervals (CIs) of each index per face gender, and control for the number of raters across the sliding windows, we randomly selected 10 raters' raw rating responses per trait from each window (which had 10 or more raters). The raw trait ratings were averaged across the 10 raters per sample for each impression per face image. Using the average ratings per face image, we then calculated (a) the interimpression correlational coefficients and (b) the amount of variance explained by PC1 in the impression ratings, a proxy for impression valence, per face gender. We repeated the rater selection and index calculation 1,000 times for each rater window as a means of bootstrapping to estimate the 95% CIs.

To understand the relationship between the rater GSE and the impression differentiation, we ran linear and quadratic regressions predicting the mean correlational coefficient and the amount of variance in male and female ratings explained by PC1 using the rater subgroup GSE score as the predictor.

The GSE score was higher than the absolute middle score of the questionnaire, 100 in the range of [20,180] (M = 121.11, SD = 17.37), t(468) = 26.33, $p < .001$, indicating that participants on average endorsed gender stereotypical beliefs (e.g., "men are more likely to be aggressive than women.", "women are more likely to be emotional than men."). Male raters showed a higher level of gender stereotype endorsement (n = 235, M = 124.71, SD = 19.64) than female raters (n = 233, M = 117.58, SD = 13.85), t(420.83) = 4.54, $p < .001$.

Importantly, when the rater GSE score increased, the correlations between impression ratings increased too for both male and female faces (Figure 3A). Although the linear regression model was significant for the ratings of both genders—male faces: R^2 = .78, F(1, 43) = 154.81, $p < .001$, and female faces: R^2 = .75, F(1, 43) = 130.63, $p < .001$—the quadratic model—male faces: R^2 = .82, F(2, 42) = 93.57, $p < .001$, and female faces: R^2 = .89, F(2, 42) = 162.07, $p < .001$—explained significantly more variance than the linear model did—male faces: F(1, 43) = 7.81, p = .008; female faces: F(1, 43) = 48.68 $p < .001$. Correspondingly, the amount of variance explained by PC1 in the ratings followed the same quadratic pattern of change across the GSE score: Although the linear regression model was significant for both genders—male faces: R^2 = .82, F(1, 43) = 191.10, $p < .001$, and female faces: R^2 = .78, F(1, 43) = 152.54, $p < .001$—the quadratic model—male faces: R^2 = .84, F(2, 42) = 110.62, $p < .001$, and female faces: R^2 = .89, F(2, 42) = 175.10, $p < .001$—explained a significantly larger amount of variance than the linear models did—male faces: F(1, 43) = 6.35, p = .016, and female faces: F(1, 43) = 44.25, $p < .001$. However, for both measures the increase was largely monotonic (see Figure 3) and the magnitude of the linear effect was larger than the magnitude of the quadratic effect.

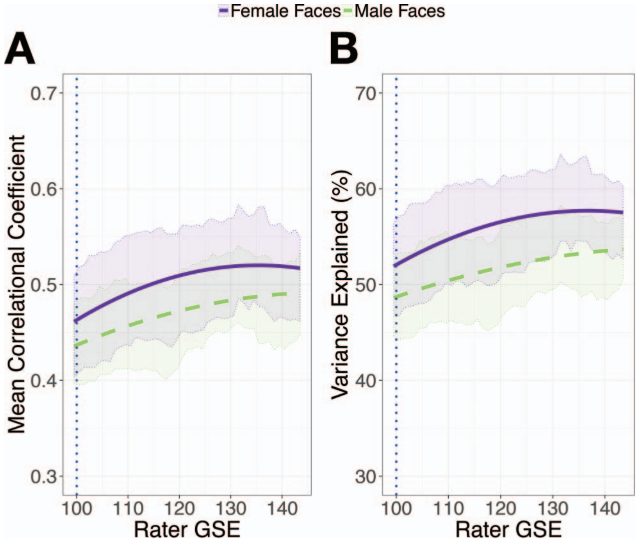

Figure 3. The level of intercorrelations across impressions (A) and the amount of variance in the impressions explained by valence (B) as a function of the raters' gender stereotype endorsement (GSE) score. Each data point (A: the absolute value of correlational coefficients between all facial impression rating pairs, B: the amount of variance explained by the first principal component (PC1) in the PCA of facial impression ratings per gender) was calculated from a rater subgroup (n_{rater} = 10 per trait, n_{rater} = 140 in total per subgroup). Each subgroup was sampled from a sliding window on the rater GSE score ($n_{\text{rater}} \geq 10$ per trait), in which the X value is the middle point of the window. The shaded regions show 95% CIs estimated from 1,000 bootstrapped replications per data point. The intercorrelations of face impressions (ts > 15.15, ps < .001) and the variance explained by PC1 were always significantly higher in female than in male impressions across the GSE score (ts > 25.70, ps < .001). The vertical dotted line at X = 100 represents no GSE bias. PCA = principal component analysis. CI = confidence interval. See the online article for the color version of this figure.

Across all rater GSE scores, female face ratings had higher correlational coefficients (Figure 3A, ts > 15.15, ps < .001) and larger amount of variance explained by PC1 than male face ratings (Figure 3B, ts > 25.70, ps < .001). We obtained consistent results when we divided the rater GSE level into four factor scores, each of which represented the level of endorsement for stereotypes about Male × Positive, Male × Negative, Female × Positive, or Female × Negative traits (Supplemental Figure S3 in the online supplemental material).

Taken together, these results show that those who more strongly endorsed gender stereotypes were more likely to form less differentiated impressions of both male and female faces. However, irrespective of this effect, raters were more likely to show less differentiated impressions of female faces than of male faces.

Role of raters' gender. To examine how the rater gender was related to differences in impression differentiation of male and female faces, we calculated correlational coefficients across impression ratings of male and female faces, separately for male and female raters. We conducted tests of matrix equality for any difference in the correlational coefficient matrices using the coefficients absolute values, across rater genders and face genders. For both male and female raters, female impressions were less differentiated than male impressions: The female face ratings were more strongly intercorrelated than male face ratings in male raters—male faces: $M_{|r|}$ = 0.50, $SD_{|r|}$ = 0.25, and female faces: $M_{|r|}$ = 0.51, $SD_{|r|}$ = 0.21; $\chi^2(91)$ = 1259.50, $p < .001$—and female raters—male faces: $M_{|r|}$ = 0.56, $SD_{|r|}$ = 0.25; female faces: $M_{|r|}$ = 0.61, $SD_{|r|}$ = 0.21; $\chi^2(91)$ = 913.18, $p < .001$. For both male and female faces, female raters showed more strongly intercorrelated impressions of faces than male raters did—male faces: $\chi^2(91)$ = 21,840.47, and female faces: $\chi^2(91)$ = 3264.53. We obtained consistent results using a 2 [face gender] × 2 [rater gender] analysis of variance (ANOVA; see online supplemental material).

To assess how the rater gender was related to the valence dependency of impressions, we conducted PCAs separately for male and female faces, this time, using the mean impression ratings of male raters and female raters. For both male and female raters, PC1 explained more variance in female than male face impressions (male raters: 55.72% vs. 54.18%; female raters: 65.69% vs. 60.77%). Taken together, these findings suggest that female raters showed less differentiated impressions of faces, especially for female than for male faces.

Study 1c: Analysis of Ma et al. (2015)

Studies 1a and 1b used the same face images. To test the robustness of the results of the previous studies, in Study 1c we run the same analyses (i.e., correlational analyses, PCAs) on a preexisting face rating dataset involving different sets of face images, impressions, and participants (Ma, Correll, & Wittenbrink, 2015).

Method.

Participants. For the impression trait ratings previously collected by Ma and colleagues (2015), over 1,087 participants (≥308 males, ≥552 females, ≥227 unreported) had rated face images on various traits. Rater gender was not included in the norming dataset and was not included in our analyses.

Stimuli. Naturalistic face photos with direct gaze and resting expressions were used. The individuals were amateur actors with no facial hair, earrings, eyeglasses, or visible make-up, all wearing gray T-shirts: 597 photos of 109 self-identified Asian (57 females), 197 Black (104 females), 108 Latino (56 females), and 183 White actors (90 females) between the ages of 17–65 were used (Ma et al., 2015). Non-White faces were included (69% of all faces), unlike Studies 1a and 1b. Prior work showed that facial evaluation space is largely common across different face races (e.g., Sutherland, Liu, et al., 2018).

Procedure. Participants rated 290 male and 307 female face photos on 15 traits—how afraid, angry, attractive, babyfaced, disgusted, dominant, feminine, happy, masculine, racially prototypical, sad, surprised, threatening, trustworthy, or unusual each individual looked. Ma et al. (2015) asked each participant to form impressions of individuals from photos on multiple attributes (e.g., "Consider the person pictured above and rate him/her with respect to other people of the same race and gender. – Fearful/Afraid (1 = Not at all; 7 = Extremely)"). As in Studies 1a and 1b, two analyses were conducted to test the gender differences in face impressions. First, to compare the in/dependency of perceived traits between genders, we contrasted the absolute values of the coefficients between male and female face ratings. We calculated pairwise correlational coefficients among all 15 trait ratings ($_{15}C_2 = 105$ pairs) for each gender. Second, to test whether raters showed less differentiated impressions for female than male faces, we calculated the amount of the variance explained by the first component for each gender.

Results and discussion. Female face impressions were, again, less differentiated and more valence-laden than male face impressions. As in Studies 1a and 1b, the correlational coefficients were significantly different, with female ratings being more strongly intercorrelated ($M_{|r|} = 0.33$, $SD_{|r|} = 0.23$) than male ratings ($M_{|r|} = 0.27$, $SD_{|r|} = 0.21$; $\chi^2(105) = 2223.24$, $p < .001$; Figure 2A, bottom), indicating a higher level of dependency between traits in female face impressions. As in Studies 1a and 1b, for both genders PC1 was highly loaded on by all positive or negative traits (Figure 1B, bottom). Correspondingly, the amount of variance explained by PC1, a proxy of valence, was larger for female than male ratings (40.87% vs. 31.60%; Figure 2B, bottom; Supplemental Table S1 in the online supplemental material), indicating a higher level of dependency of female ratings on valence. The 15 traits used in Study 1c were different from the 14 traits rated in Studies 1a and 1b. The traits in Studies 1a and 1b were determined based on the frequency of mention in unconstrained verbal descriptions of face images. This difference may explain the small difference in the results between the studies (e.g., overall lower level of intertrait correlation in Study 1c). All in all, the results in Study 1c replicate what we found in Studies 1a and 1b: Face impressions of women are less differentiated and are more highly valence-laden than those of men.

Studies 2 and 3: Computational Models of Male and Female Face Impressions

Study 2: Building Gender-Specific Impression Models

Study 1 showed that face impressions of women are less differentiated and are more highly valence-laden than those of men. Studies 2 and 3 examine the basis of this phenomenon: How is the gender difference in impressions related to people's use of facial information when forming impressions (e.g., using facial masculinity to form an impression of dominance) of men and women? Does the gender difference stem from (a) earlier, lower-level, differences in perception of male and female faces or (b) later, higher-level, evaluative differences? Specifically, people may either (a) use different facial information when forming impressions of men and women or (2) use the same facial information when forming impressions of both men and women but evaluate this information differently. We built and validated impression models to test these possibilities. We model impressions for each gender rather than modeling them collapsed across genders as in previous work (Oosterhof & Todorov, 2008).

To investigate the extent to which people use dis/similar facial information to form impressions of men and women, we built data-driven models of impressions for male and female faces and calculated the similarities between these gender-specific face impression models. Data-driven face modeling reveals facial information that correlates with an impression with little prior assumptions of what information matters (Funk, Walker, & Todorov, 2017; Jack & Schyns, 2017; Oosterhof & Todorov, 2008; Todorov et al., 2011; Todorov & Oosterhof, 2011; Walker & Vetter, 2009, 2016). In prior work (Oosterhof & Todorov, 2008), participants rated randomly generated faces on a trait (e.g., how competent they looked). The faces were generated from a multidimensional, statistical face space, where each face is represented as a point in this space. In this approach, each impression model is a vector in the space that visualizes holistic changes in facial appearance that make the face appear more trait-like (e.g., more competent). All prior computational models were built irrespective of the gender of faces. Here, we built models of impressions of trustworthiness and dominance separately for male and female faces.

Method.

Participants. Five-hundred-and-ten MTurk online workers living in the United States (233 males, 256 females, 21 unreported) participated for monetary reward.

Stimuli and procedure. Previous data-driven computational models of face impressions were based on trait impression ratings of randomly generated faces from a multidimensional, statistical face space (Oosterhof & Todorov, 2008). To build new models separately for male and female faces, we generated 301 male faces and 301 female faces with FaceGen 3.2 (Singular Inversions, Toronto, Canada). The FaceGen model is based on a database of actual male and female faces that were laser-scanned in 3D. These sample faces consist of individuals of diverse races (e.g., East Asian, Black, West Asian, White). In the FaceGen model, a face is a point in a 100-dimensional face space. The 100 dimensions are orthogonal to each other and are chosen to capture large variance in the appearance of individual faces. Moving a face along a dimension in this space results in a holistic change in the shape and reflectance (i.e., texture and coloration) of the face in a specific way. The shape and the reflectance of a face are determined by 50 shape and 50 reflectance parameters. We generated male and female faces (300 each) by randomly selecting each parameter from a normal distribution. We used a single set of 300 source faces and made them more male- or female-like, so that all the male and all the female faces were centered around the average male and the average female faces in the FaceGen database, respectively. This resulted in paired images of male and female faces (Supplemental Figure S4 in the online supplemental material). All stimuli are available at Open Science Framework: https://osf.io/ycv72/.

Each participant rated either 51 male or 51 female faces on one of two impressions—trustworthiness or dominance. We chose these two impressions, because they are the best approximations of the valence and power dimensions underlying face impressions, and they are highly distinctive from each other (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Todorov, Said, Engell, & Oosterhof, 2008). We kept the same participants from participating in both "trustworthiness" and "dominance" rating tasks. The random 51 faces of each gender consisted of 50 random faces from the pool of 300 random faces and the gender-specific average face. Participants were told that there were no right or wrong answers and that we were interested in their first impression or "gut response." The faces were presented twice in two separate blocks for the reduction of the measurement error and the screening of unreliable raters' data. We calculated each rater's intrarater reliability by correlating their ratings from the two different blocks. The ratings from participants with zero or negative intrarater reliability were excluded, which left us with responses of 418 participants (181 males, 217 females, and 20 unreported). Each face was presented in color (512 × 512 pixels with the height of the face being about 440 pixels) with a question below it, "How [*trait term*, e.g., trustworthy] is this [man/woman]?" ranging from 1 (Not at all *[trait term]*) to 9 (Extremely *[trait term]*). Each face was visible until the participant responded, the ITI was 250 ms, and the order of faces was randomized. The ratings of trustworthiness and dominance were negatively correlated for both genders, but the correlation was stronger for female faces (Supplemental Figure S5 in the online supplemental material).

To create gender-specific computational trait models—male trustworthiness model, female trustworthiness model, male dominance model, and female dominance model—we averaged the two ratings per face across participants for each trait. For each gender-specific model, we computed the contribution of each of the 50 shape and the 50 reflectance parameters to the average trait impression ratings of the 300 faces, following the previous data-driven statistical approach (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011). The mean impression ratings of the 301 faces and the values for a single parameter (out of 100) of the 301 faces are essentially two vectors with 301 elements each. To create one parameter of an impression model, the cross-product of these two vectors were summed across faces, and then were normalized across parameters. The 100 parameters of the model represented the amount of variation that would induce a 1SD change in the trait impression rating.

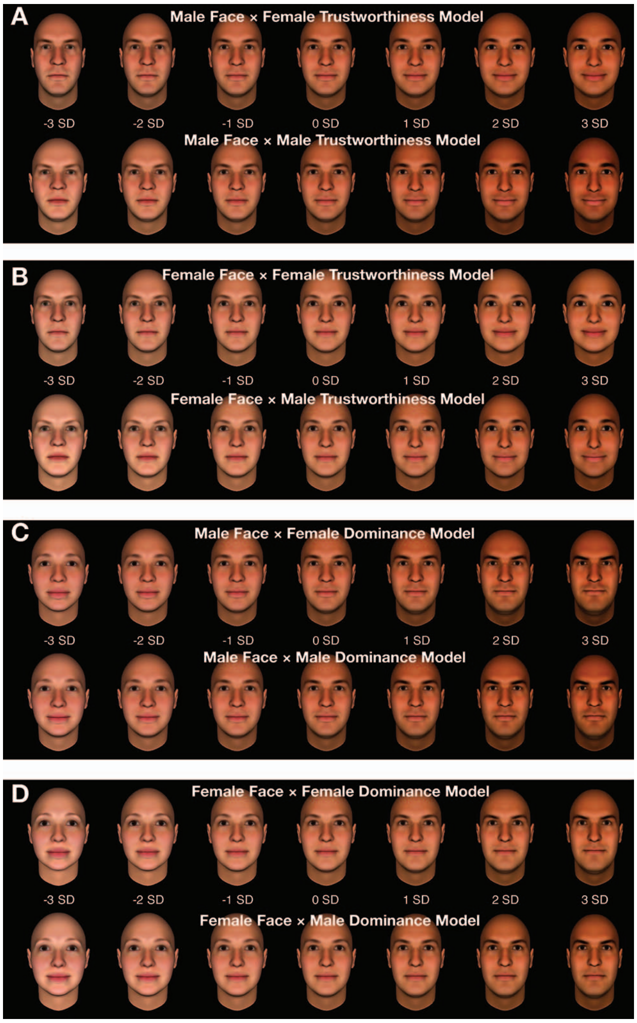
A
Male Face × Female Trustworthiness Model
-3 SD
-2 SD
-1 SD
0 SD
1 SD
2 SD
3 SD
Male Face × Male Trustworthiness Model
B
Female Face × Female Trustworthiness Model
-3 SD
-2 SD
-1 SD
0 SD
1 SD
2 SD
3 SD
Female Face × Male Trustworthiness Model
C
Male Face × Female Dominance Model
-3 SD
-2 SD
-1 SD
0 SD
1 SD
2 SD
3 SD
Male Face × Male Dominance Model
D
Female Face × Female Dominance Model
-3 SD
-2 SD
-1 SD
0 SD
1 SD
2 SD
3 SD
Female Face × Male Dominance Model

Figure 4. Gender-specific models of facial impressions of trustworthiness and dominance applied to novel synthetic male and female face images in Study 3a. The trustworthiness (A, B) and dominance impression models (C, D) derived from ratings of female (the top row of each subpanel) and male faces (the bottom row of each subpanel) were applied to a sample male (A, C) and female face (B, D). Both male and female models could manipulate the impression of both male and female faces, showing that facial information used to form these key impressions is similar across face genders. See the online article for the color version of this figure.

Results and discussion. The resulting gender-specific statistical impression models are shown in Figure 4. Both models derived from male and female face ratings are similar to existing statistical models of impressions (e.g., Oosterhof & Todorov, 2008). Further, the gender-specific models represent similar facial information found in prior research: As both male and female faces are manipulated to appear more trustworthy, their expressions become more positive and vice versa (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009). Likewise, as both male and female faces are manipulated to appear more dominant, they become more masculine and facially mature (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Zebrowitz, 2004; Zebrowitz & Montepare, 2008).

Within each impression, male and female models were highly positively correlated, suggesting that similar information is used when people form impressions of male and female faces (trustworthiness: ρ = .68, dominance: ρ = .85; Supplemental Figure S6 and Supplemental Table S3 in the online supplemental material). However, trustworthiness and dominance models were more strongly negatively correlated in the female (ρ = −.38) than in the male models (ρ = −.16), suggesting that these models are more similar for female than for male faces. This is consistent with the data from Studies 1a–1c, in which face impressions of women were less differentiated than those of men.

Role of raters' gender. The correlation between the female trustworthiness and dominance models was stronger for female (ρ = −.42) than male raters (ρ = −.30), whereas the correlation between the male trustworthiness and dominance models was comparable for female (ρ = −.16) and male raters (ρ = −.16). These findings show that female raters relied on more similar visual information when forming impressions of trustworthiness and dominance of female faces than did male raters. This is consistent with the data from Study 1b, in which female raters showed less differentiated impressions of female than of male faces.

Study 3a: Cross-Gender Validation of Models With Synthetic Face Images

In Study 3, we cross-validated the gender-specific impression models. Specifically, we manipulated the level of perceived traits of novel images of male and female faces (Supplemental Figure S7 in the online supplemental material) and asked a new group of participants to rate the faces on the respective traits. The first objective of this study was to test whether the models of impressions successfully capture the changes in facial appearance that lead to changes in impressions. The second, more important objective was to test whether the gender-specific impression models work better when applied to a congruent face (e.g., when a male model is applied to a male face). We test two possible outcomes in Studies 3a and 3b.

If the face models apply better to the gender-congruent faces (e.g., male models apply better to male than female faces) despite the similarity of the gender-specific models observed in Study 2, then it would suggest that the gender differences in impressions are likely due to lower-level, perceptual (rather than later, evaluative) differences in the usage of visual cues when forming facial impressions of men and women.

In contrast, if the face models apply equally well to male and female faces (e.g., male models apply equally well to male and female faces), it would suggest that the gender differences in impressions are likely due to later, evaluative (rather than lower-level, perceptual) differences when forming facial impressions of men and women. Such a result would suggest that people use similar visual information when forming impressions of male and female faces, although they interpret this information differently (e.g., whereas a masculine male face is evaluated positively, a masculine female face is evaluated negatively; Sutherland et al., 2015; Oh et al., 2019).

Method.

Participants. Two hundred and sixty-eight MTurk online workers living in the United States (121 males, 147 females) participated for monetary reward.

Stimuli and procedure. To validate the gender-specific face models, we generated faces that reflected the impression change in each model. First, using a procedure similar to previous validation studies (Todorov et al., 2013), we generated 25 new faces: We generated 1,000 faces whose positions on the 100 parameters were independently sampled from 100 normal distributions. From the 1,000 random faces, 25 faces that physically differed maximally to one another were chosen (i.e., faces with highest average euclidean distance to each other; Supplemental Figure S7 in the online supplemental material). We used a single set of 25 faces and made them more male- or female-like. All stimuli are available at Open Science Framework: https://osf.io/ycv72/.

Second, we manipulated each face with the trait models. We varied the face parameters of the 50 faces by adding −3, −2, −1, 0, 1, 2, and 3 *SD*s, with the 0 *SD* addition being null manipulation (see the online supplemental material). There were four gender-specific models—male trustworthiness, female trustworthiness, male dominance, and female dominance models. This resulted in 1,400 faces (2 [face gender] × 25 [identities] × 7 [manipulation levels] × 4 [model]).

Each participant rated either male or female faces manipulated by either male or female model of either trustworthiness or dominance. We kept the same participants from participating in more than one task (e.g., participating in both the "male trustworthiness" rating and the "female dominance" rating tasks). The face stimuli were presented in color on the screen (512 × 512 pixels with the height of the face being about 440 pixels). Participants were told that there were no right or wrong answers and that we were interested in their first impression: 175 faces were presented in a random order (25 [identities] × 7 [levels]), followed by presentation of 25 randomly chosen faces from the previously presented faces without any noticeable break. The 25 faces were repeated for the calculation of test–retest reliability. The ratings from participants with zero or negative intrarater reliability were excluded, which left us with ratings of 247 raters (107 males, 140 females). Each face Gender × Model Trait × Model Gender condition had ≥30 raters. The ratings of trustworthiness and dominance were highly reliable irrespective of whether the gender of the original faces and the model were identical ($\alpha_{\min}$ = .96) or not ($\alpha_{\min}$ = .96; Table S4).

To assess how well the models varied the intended impressions, we ran linear and quadratic regressions for the trait ratings of male and female faces with the level of model manipulation as the predictor. To determine whether the gender-specific models are more successful when applied to a congruent face (e.g., a female model applied to a female face), we then compared the predictive powers of the models across genders. We ran repeated-measures ANOVAs on Fisher's *z* scores, transformed from the correlations between the observed and predicted ratings of the regressions.

Results and discussion. All linear and quadratic models explained significant amount of variance in the impression ratings regardless of whether the model gender and the face gender were congruent (linear: mean R^2s > .94, quadratic: mean R^2s > .96; Figure 5) or not (linear: mean R^2s > .94, quadratic: mean R^2s > .97). Thus, the effectiveness of the trait model was not affected by the congruency between the face gender and the model gender.

To further assess the relative effectiveness of the models, we ran a repeated measures ANOVA on the correlations between the predicted and observed ratings, transformed to Fisher's *z* scores, for both the linear and quadratic regression models. We found a

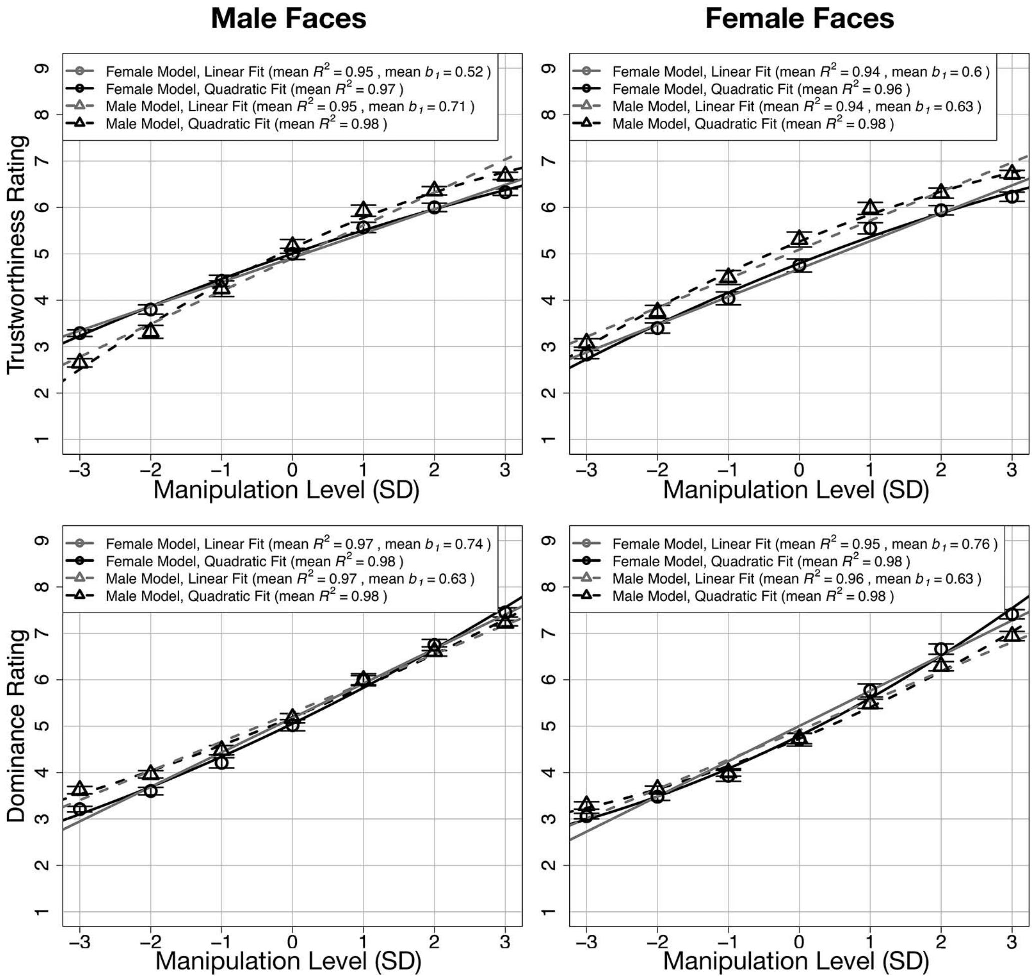

Figure 5. Validation of models of trustworthiness (top) and dominance (bottom) with synthetic male (left) and female faces (right). Linear (gray) and quadratic (black) fit of ratings of trustworthiness as a function of the female (solid) and male (dashed) model values of the faces. The mean coefficient of determination (R^2) and unstandardized coefficient (b_1) averaged across faces per model are displayed. Error bars denote ± *SE*.

main effect of model trait for the linear, $F(1, 24) = 5.19$, $p = .03$, $\eta^2_G = .05$, and the quadratic regression models, $F(1, 24) = 5.98$, $p = .02$, $\eta^2_G = .03$. Specifically, the dominance models predicted the ratings better (linear: $M_z = 2.42$, $SD_z = 0.36$; quadratic: $M_z = 2.75$, $SD_z = 0.42$) than the trustworthiness models did (linear: $M_z = 2.26$, $SD_z = 0.35$; quadratic: $M_z = 2.63$, $SD_z = 0.35$). We also found a main effect of face gender for the linear, $F(1, 24) = 10.52$, $p < .01$, $\eta^2_G = .04$, and quadratic regression models, $F(1, 24) = 4.40$, $p < .05$, $\eta^2_G = .02$. Specifically, the ratings of male faces were better predicted by the models (linear: $M_z = 2.41$, $SD_z = 0.37$; quadratic: $M_z = 2.74$, $SD_z = 0.40$) than the ratings of female faces (linear: $M_z = 2.28$, $SD_z = 0.35$; quadratic: $M_z = 2.65$, $SD_z = 0.38$). These two effects were not predicted and were relatively small in size. Only for the quadratic regression model, we found a significant Model Gender × Face Gender interaction, $F(1, 24) = 7.09$, $P = .01$, $\eta^2_G = .03$, indicating that the male models were better at manipulating impressions of female faces ($M_z = 2.75$, $SD_z = 0.42$) than male faces ($M_z = 2.71$, $SD_z = 0.32$), whereas the female models were better at manipulating impressions of male faces ($M_z = 2.76$, $SD_z = 0.47$) than female faces ($M_z = 2.54$, $SD_z = 0.32$).

In sum, the models could generate faces varying on the intended impressions within and across gender. The effect of gender congruency was found only in the quadratic models: The direction of the effect was the opposite of what was hypothesized, and the effect size was relatively small. We conducted another validation study to test the robustness of these effects.

Study 3b: Cross-Gender Validation of Models With Real-Life Face Images

In Study 3a, we manipulated male and female faces to look more/less trustworthy or dominant using male and female impression models and calculated how well the models predicted the ratings of gender-congruent and incongruent faces. We found no evidence for gender specificity in the models' capacity to manipulate impressions. If anything, for the quadratic models, the predicted interaction was in the opposite direction. These findings suggest that people may be using similar visual information when forming impressions of male and female faces. To assess the robustness of these results, in Study 3b, we conducted another validation study using real-life face images. Specifically, using the gender-specific models, we manipulated the level of perceived traits of novel real-life images of male and female faces. We then asked a new group of participants to rate the faces on the respective traits.

Method.

Participants. Two hundred and ninety-two MTurk online workers living in the United States (127 males, 164 females, and one other gender) participated for monetary reward.

Stimuli and procedure. Photos of male and female faces (25 each) were randomly selected from a face database (DeBruine & Jones, 2017), consisting of naturalistic face photos with direct gaze and resting expressions without any eyeglasses or visible make-up, all wearing white T-shirts: 50 photos of eight self-identified East Asian (four females), 8 West Asian (three females), 12 Black (five females), and 22 White actors (13 females) between the ages of 19–37 were used (Supplemental Figure S8 in the online supplemental material). Non-White faces comprised 56% of all faces. The face-space model on which the gender-specific models are built is based on the 3D scans of White and non-White faces, and the face-space model can capture the trait-related facial variance within and across different facial races. All stimuli are available at Open Science Framework: https://osf.io/ycv72/.

Using the impression models, we manipulated the faces along the respective traits. As in Study 3a, we prepared seven facial variations, including the original face, for each Face Identity × Impression Model. To apply the models, we transformed the initial face images along the gender-specific impression models from Study 2, using PsychoMorph (Tiddeman, Burt, & Perrett, 2001). First, we created synthetic faces that represented these models, extreme faces that are −4 and 4SD deviant from the average FaceGen face on each model. Next, we used the transformation function of PsychoMorph to change the initial 25 male and 25 female real-life face images (Supplemental Figure S8 in the online supplemental material) along the continuum of the difference between the two extreme face images, for each impression. Unlike the standard morphing procedure, which is a direct transition between two images, the transformation procedure in PsychoMorph allows users to manipulate a single image along a continuum and generate photo-realistic images (Sutherland, Rhodes, & Young, 2017). On the most extreme ends, each face image was transformed 40% toward the −4 or 4 *SD* model face. The value of 40% was chosen because any stronger manipulation caused distortional artifacts on the images. The manipulation magnitude was identical for the intervals between the seven facial variations: The final face images were transformed −40%, −26.67%, −13.33%, 0%, 13.33%, 26.67%, and 40% from the initial faces (see Figure 6). To maintain the ostensible gender and ethnicities of the original faces, we only used the variation in the face shape of the models. As in Study 3a, there were four gender-specific models. This resulted in 1,400 faces (2 [face gender] × 25 [identities] × 7 [manipulation levels] × 4 [model]).

As in Study 3a, each participant rated either male or female faces manipulated by either male or female model of either trustworthiness or dominance. We kept the same participants from participating in more than one rating task. The face stimuli were presented in color (512 × 512 pixels with the height of the face being about 440 pixels). The rating design was identical with Study 3a: 175 faces were presented first (25 [identities] × 7 [levels]), followed by presentation of 25 randomly chosen faces from the previously presented faces without any break. The 25 faces were repeated for the calculation of test–retest reliability. The ratings from participants with zero or negative intrarater reliability were excluded, which left us with ratings of 239 raters (138 males, 100 females, and 1 other gender). Each Face Gender × Model Trait × Model Gender condition had ≥30 raters. The ratings of both trustworthiness and dominance were reliable irre-

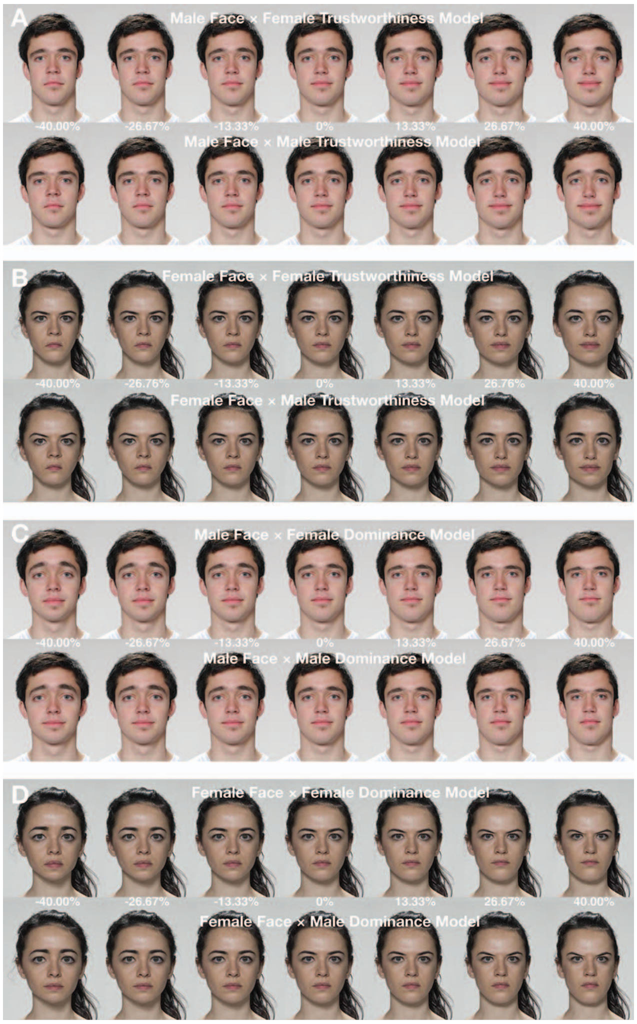
A
Male Face × Female Trustworthiness Model
-40.00%
-26.67%
-13.33%
0%
13.33%
26.67%
40.00%
Male Face × Male Trustworthiness Model
B
Female Face × Female Trustworthiness Model
-40.00%
-26.76%
-13.33%
0%
13.33%
26.76%
40.00%
Female Face × Male Trustworthiness Model
C
Male Face × Female Dominance Model
-40.00%
-26.67%
-13.33%
0%
13.33%
26.67%
40.00%
Male Face × Male Dominance Model
D
Female Face × Female Dominance Model
-40.00%
-26.67%
-13.33%
0%
13.33%
26.67%
40.00%
Female Face × Male Dominance Model

Figure 6. The gender-specific models of facial impressions of trustworthiness and dominance applied to real-life face images in Study 3b. The trustworthiness (A, B) and dominance impression models (C, D) derived from ratings of female (the top row of each subpanel) and male faces (the bottom row of each subpanel) were applied to a sample male (A, C) and female faces (B, D). Both male and female models could manipulate the impression of both male and female faces, showing that facial information used to form these key impressions is similar across face genders. Images from *Face research lab London set*, by L. M. DeBruine and B. C. Jones, 2017. Retrieved from http://dx.doi.org/10.6084/m9.figshare.5047666. See the online article for the color version of this figure.

spective of whether the gender of the original faces and the model were identical ($\alpha_{\min} = .88$) or not ($\alpha_{\min} = .83$; Table S5).

To assess how well the models varied the intended impressions of faces, we adopted the same analyses as those in Study 3a—regressions for the trait ratings and repeated measures ANOVAs on Fisher's *z* scores converted from the correlations between the observed and predicted ratings of the regressions.

Results and discussion. As in Study 3a, all models significantly explained the impression ratings regardless of whether the model gender and the face gender were congruent (linear: mean R^2s > .54, quadratic: mean R^2s > .68; Figure 7) or not (linear: mean R^2s > .63, quadratic: mean R^2s > .68). To further assess the relative effectiveness of the models, we ran a 2 [face gender] × 2 [model trait] × 2 [model gender] repeated measures ANOVA on Fisher's *z* scores converted from the correlations between the predicted and observed ratings for both the linear and quadratic regression models.

We found a significant main effect of model trait for the linear, $F(1, 24) = 82.73$, $p < .001$, $\eta^2_G = .35$, and the quadratic regression models, $F(1, 24) = 75.49$, $p < .001$, $\eta^2_G = .33$. Specifically, the dominance models predicted the ratings better (linear: $M_z = 1.87$, $SD_z = 0.45$; quadratic: $M_z = 2.11$, $SD_z = 0.48$) than the trustworthiness models did (linear: $M_z = 1.22$, $SD_z = 0.49$; quadratic: $M_z = 1.46$, $SD_z = 0.49$). The same effect was observed in Study 3a.

We also found a main effect of model gender for the linear, $F(1, 24) = 4.67$, $p < .05$, $\eta^2_G = .02$, and the quadratic regression models, $F(1, 24) = 4.93$, $p = .04$, $\eta^2_G = .02$. Specifically, models derived from ratings of male faces were more effective in manipulating impressions (linear: $M_z = 1.61$, $SD_z = 0.52$; quadratic: $M_z = 1.86$, $SD_z = 0.51$) than models derived from ratings of female faces (linear: $M_z = 1.49$, $SD_z = 0.61$; quadratic: $M_z = 1.71$, $SD_z = 0.64$). We also found a significant Model Trait × Face Gender interaction for the linear, $F(1, 24) = 7.84$, $p < .01$, $\eta^2_G = .03$, and the quadratic regression models, $F(1, 24) = 8.66$, $p < .01$, $\eta^2_G = .03$, indicating that dominance models were more effective at manipulating female (linear: $M_z = 1.91$, $SD_z = 0.50$; quadratic: $M_z = 2.22$, $SD_z = 0.52$) than male faces (linear: $M_z = 1.84$, $SD_z =$

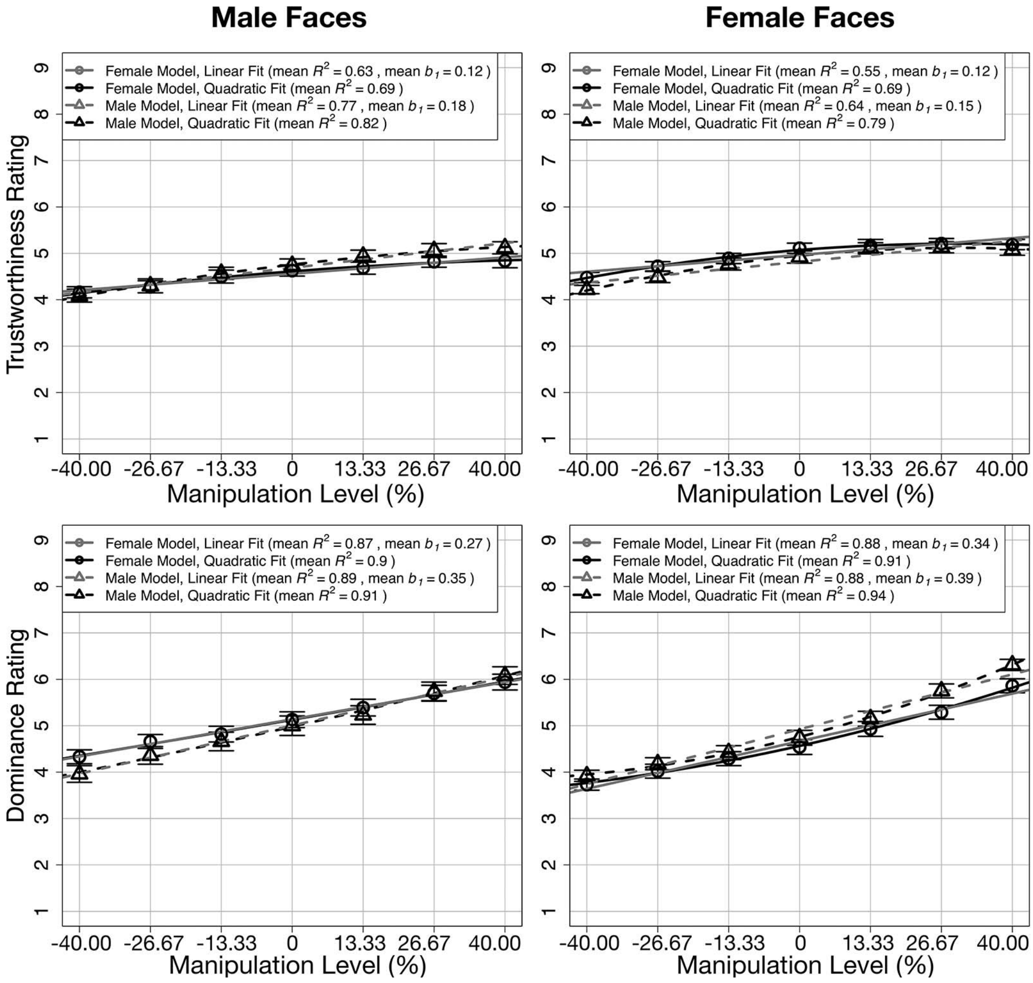

Figure 7. Validation of models of trustworthiness (top) and dominance (bottom) with real-life male (left) and female face images (right). Linear (gray) and quadratic (black) fit of ratings of trustworthiness as a function of the female (solid) and male (dashed) model values of the faces. The mean coefficient of determination (R^2) and unstandardized coefficient (b_1) averaged across faces per model are displayed. Error bars denote ± *SE*.

0.39; quadratic: $M_z = 2.00$, $SD_z = 0.41$), whereas the trustworthiness models were more effective at manipulating male (linear: $M_z = 1.34$, $SD_z = 0.49$; quadratic: $M_z = 1.51$, $SD_z = 0.53$) than female faces (linear: $M_z = 1.09$, $SD_z = 0.45$; quadratic: $M_z = 1.42$, $SD_z = 0.46$). None of these two effects were observed in Study 3a.

Across the two validation studies with computer-generated and real-life faces, the only consistent finding was that the dominance impression models could generate faces varying on the intended impressions better than the trustworthiness impression models could, irrespective of the face gender or the model gender. All in all, the models were capable of generating faces varying on the intended impressions within and across gender, showing no evidence for gender specificity.

General Discussion

People form appearance-based impressions of men and women differently. Women whose faces appear more dominant, for example, are perceived as less trustworthy, whereas the perceived dominance of men does not affect impressions of their trustworthiness (Sutherland et al., 2015; Sutherland, Liu, et al., 2018). In the current series of studies, we addressed two questions: Are impressions of women less differentiated than impressions of men in general and what is the source of this difference in impressions? Answering these questions is important, because simplified impressions of women can lead to unfair treatment: Women are evaluated negatively when they behave not “woman-like” (e.g., assertive; Moss-Racusin et al., 2012; Rudman, 1998; Rudman & Phelan, 2008).

By analyzing multiple dataset of facial impressions and building separate face models of impressions of men and women, we found that (a) first impressions from faces are less differentiated for women and importantly that (b) visual information used to form impressions of trustworthiness and dominance is highly similar for men’s and women’s faces. Combined, these findings suggest that people use the same visual information when forming impressions of men and women, but that this information is evaluated differently. These results emphasize the role of gender categorization in impression formation. The present article also shows that trait impressions from faces are more strongly intercorrelated in those with stronger stereotypes. Finally, the present article introduces the first separate data-driven impression models for male and female faces.

Our first main finding was that participants held less differentiated (i.e., more highly intercorrelated) trait impressions of women than of men (Studies 1–2). Specifically, trait impressions of women varied from each other to a smaller degree and were more tied to overall positivity/negativity evaluation than impressions of men (Study 1). Consistent with this finding, the models of trustworthiness and dominance impressions were more similar for female than for male faces (Study 2). These findings confirm ambivalent sexism (Glick & Fiske, 1996, 2011) in the domain of visual perception, corroborating previous research (Sutherland et al., 2015). The theory of ambivalent sexism posits that women are evaluated positively as far as they are stereotyped into restricted traits and roles (e.g., being helpful), unlike men whose impression valence is less dependent on stereotypes. This theory would predict that women whose appearance evokes counterstereotypical trait impressions (e.g., assertiveness) are likely to be evaluated more negatively than men whose appearance evokes counterstereotypical trait impressions (e.g., tenderness). This gender difference will lead to a higher level of intercorrelations between the impressions of women than between those of men. In the same vein, our findings suggest that the backlash effect, a phenomenon in which women who violate prescriptions of feminine traits receive social/economic penalties (Rudman, 1998; Rudman & Glick, 2001; Rudman & Phelan, 2008), generalizes to visual perception. Qualities perceived as traditionally masculine lead to more negative impressions of women (e.g., less likable): Women with dominant facial looks (Sutherland et al., 2015), assertive attitude (Rudman, 1998; Rudman & Glick, 2001; Rudman & Phelan, 2008), or work competency (Hagen & Kahn, 1975) are evaluated more negatively than women with the opposite qualities. Although qualities perceived as feminine usually lead to more negative impressions of men (Derlega & Chaikin, 1976; Heilman & Wallen, 2010; Moss-Racusin, Phelan, & Rudman, 2010), the effects of counterstereotypical looks on men are weaker (Sutherland et al., 2015) and sometimes even beneficial. Men (and women) with a feminine face shape, for example, are perceived as attractive (Perrett et al., 1998; Rhodes, Hickford, & Jeffery, 2000; Said & Todorov, 2011; but see Rhodes, 2006). In sum, when gender norms are violated, women face harsher penalties than men. Because people judge traits from faces, including norm-related traits, and face evaluation broadly depends on the valence of the impression, the stronger negative consequences of norm violation for women can lead to stronger links between multiple trait judgments.

The high level of dependency on valence (the first principal component) and the weak dependency on the second component in female impressions highlights an overall low level of dependency of impressions on dominance for female faces (Figure 2 and Supplemental Table S1 in the online supplemental material). Although a smaller variance in dominance ratings in female faces could explain the low dominance dependency, dehumanization of women might also underlie the phenomenon. A person is perceived as less agentic and powerful, thus less dominant, when they are perceived as more object-like (Haslam, 2006). Women are often perceived as more object-like (i.e., lacking autonomy) than men are (Nussbaum, 1999). The dehumanization account would also explain the simpler trait structure in impressions of female faces.

The level of impression differentiation was affected by the characteristics of those who formed the impressions. Raters who were more likely to endorse gender stereotypes showed less differentiated impressions of both men and women than raters who were less likely to endorse these stereotypes. This finding supports the idea that gender differences in impression differentiation result from evaluative processes triggered by gender categorization and, more generally, the idea that trait impressions of a group can be more strongly or weakly intercorrelated depending on the perceiver’s stereotypes of the group’s traits (Secord & Berscheid, 1963; Stolier, Hehman, & Freeman, 2018). Those who strongly expect gender-stereotypic qualities in others would show a smaller within-gender variation and a bigger between-gender variation in impressions of others than those who do not. This could lead to stronger intercorrelations between impressions in those who strongly endorse gender stereotypes. These individuals may evaluate others based on their gender category, because gender stereotypes are highly accessible to them during impression formation (Higgins, 1996; Lepore & Brown, 1997).

Participant gender also affected the level of impression differentiation. Surprisingly, it was female raters who showed less differentiated impressions of women (Study 1b). Correspondingly, models of trustworthiness and dominance impressions derived from female ratings were also more similar than the models derived from male ratings (Study 2). This participant gender effect may seem surprising, because female participants are less likely to endorse gender stereotypes or traditional sex roles than male participants (Study 1b; Glick & Fiske, 1996; Swim et al., 1995; Williams & Best, 1990). However, female participants have previously been found to show person evaluation consistent with gender stereotypes (e.g., Garcia-Retamero & López-Zafra, 2006; Goldberg, 1968; Parks-Stamm et al., 2008; Rudman, 1998). In the case of visual person impressions, female participants' more simplified female impressions might arise because of their sensitivity to the a/typicality of female faces. Female raters might have a clearer prototype of female faces than male raters and therefore might be more sensitive to the typicality in female faces. Given that facial typicality affects impression valence (Dotsch, Hassin, & Todorov, 2016; Sofer, Dotsch, Wigboldus, & Todorov, 2015), this may explain why female participants showed lower levels of impression differentiation and higher levels of valence-dependency for female (vs. male) faces. Bolstering such a possibility is women's better recognition of female faces, relative to men (Ellis, Shepherd, & Bruce, 1973; Lewin & Herlitz, 2002; Rehnman & Herlitz, 2006). Women have been repeatedly found to be better at recognizing female faces than men, whereas there have been little cross-gender observer differences in recognizing male faces. This specific female-on-female recognition superiority suggests that women may be better at noticing facial differences across individual females although this ability could be based on other mechanisms (e.g., better encoding of other females' faces). If this ability of female raters contributes to their higher level of sensitivity to facial typicality of female faces, this could lead to their lower impression differentiation level and higher level of valence-dependency for female faces.

Our second main finding was that the models of trustworthiness and dominance impressions of male and female faces were based on similar facial information (Studies 2 and 3). These are the first computational models of trustworthiness and dominance impressions built separately for men and women. Correlational analyses (Study 2) and cross-gender validations of these models (Study 3) consistently found that similar facial information is used in these key impressions of both genders. Specifically, the model visualizations (Figures 4 and 6) show that resemblance to emotional expressions is a key input to trustworthiness impressions (Engell, Todorov, & Haxby, 2010; Hess, Blairy, & Kleck, 2000; Keating, Mazur, & Segall, 1981; Montepare & Dobish, 2003; Oosterhof & Todorov, 2008, 2009; Said, Sebe, & Todorov, 2009; Sutherland et al., 2013; Zebrowitz & Montepare, 2008) for both male- and female-based models. Consistent with prior models (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009), as the faces are manipulated to look more trustworthy, they acquire more positive expressions. In contrast, as the faces are manipulated to look less trustworthy, they acquire more negative expressions. For dominance impressions, the key inputs are masculinity and facial maturity (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Zebrowitz, 2004; Zebrowitz & Montepare, 2008) and to smaller extent similarity to angry facial expressions (Hareli, Shomrat, & Hess, 2009; Hess et al., 2000; Said et al., 2009). As the faces are manipulated to look more dominant, they become more masculine, facially mature, and acquire more negative expressions. In contrast, as the faces are manipulated to look less dominant, they become more feminine, babyfaced, and acquire more positive expressions.

Comparing the male- and female-based models, it is also possible to see specific gender differences, especially for the models of trustworthiness impressions. For example, faces on the positive end of the female trustworthiness model are more light-skinned than faces on the positive end of the male trustworthiness model (Figures 4 and 6), possibly reflecting real gender differences in the human face (Jablonski & Chaplin, 2000). However, as the cross-validation results showed, these gender differences did not matter for impressions. Impressions of male and female faces were successfully manipulated irrespective of whether the model was derived from ratings of female or male faces, and this was the case for both synthetic and real-life faces (Study 3).[2] These findings clearly show that people use highly similar information when forming impressions of men and women on trustworthiness and dominance, but they evaluate this information differently.

The high similarity in facial impression models is consistent with the findings of South Palomares, Sutherland, and Young (2018), who built male- and female-specific face models that represent traits preferred in a romantic partner, and found that the two models are highly similar. By showing that there is little low-level perceptual differences in the information people use to form impressions of male and female faces, the current findings highlight the importance of social categorization and the perceiver's preconceptions about categories in person impressions (Fiske & Neuberg, 1990; Freeman, Penner, Saperstein, Scheutz, & Ambady, 2011; Hugenberg & Bodenhausen, 2004; Kramer, Young, Day, & Burton, 2017; Stolier & Freeman, 2016, 2017).

Our conclusion that people evaluate information differently from male and female faces (rather than use different information for male and female faces) is inferred indirectly by combining the results of multiple studies: Despite the higher level of intercorrelations and valence-dependency in women's impressions (Studies 1 and 2), both male- and female-specific impression models could manipulate facial impressions irrespective of the gender of the face (Study 3); further, the gender stereotype endorsement of participants was predictive of the level of facial impression differentiation (Study 1b). Future research can test this conclusion in a more direct fashion. One, for example, can measure the stereotype endorsement level of participants whose ratings are used to build gender-specific models: Female impression models derived from those who strongly endorse gender stereotypes would be more similar to each other across traits than female models derived from those who do not endorse gender stereotypes.

The gender difference in the structure of impressions has implications for both social perception theories and social justice. Although often visually ambiguous and conceptually continuous,

[2] Although we did not observe differences in facial information used to form impressions of male and female faces, it is of course possible that more sensitive models could reveal such differences.

gender is thought as categorical (Fiske & Neuberg, 1990) and people judge the gender of faces with ease and a high level of consensus (Hehman, Sutherland, Flake, & Slepian, 2017). Moreover, gender-related differences in facial features are easily detectable from faces (Burriss, Little, & Nelson, 2007; Schyns, Bonnar, & Gosselin, 2002) and are processed in the early stages of face perception (Cellerino et al., 2007; Mouchetant-Rostaing, Giard, Bentin, Aguera, & Pernier, 2000; Mouchetant-Rostaing & Giard, 2003; Welling, Bestelmeyer, Jones, DeBruine, & Allan, 2017). Further, facial information that correlates with gender (e.g., facial masculinity) shapes the formation of person impressions (Oh et al., 2019; Oosterhof & Todorov, 2008; Sutherland et al., 2013, 2015).

For social perception theories, our findings suggest that gender categorization shapes the overall pattern of person impression formation over and beyond associations between a few facial features and trait impressions. Beyond the differences in impressions of male and female faces, our methods can be extended to test for similar effects for other meaningful face subcategories in which one or more subcategories are more stereotyped or prejudiced against than the others (e.g., race and ethnicity). For instance, impressions of Black individuals might be less differentiated and more valence-laden than impressions of White individuals when rated by individuals living in the United States.

Further research could identify to what extent one could attenuate or reverse the gender difference in facial evaluation induced by the face category. Although we used static facial images here, nonstatic facial cues (e.g., dynamic facial gestures; Gill, Garrod, Jack, & Schyns, 2014) and nonfacial cues (e.g., clothes; Freeman et al., 2011; Oh, Shafir, & Todorov, 2019) strongly affect social perception. Gill et al. (2014), for example, found that trait-related facial movements (e.g., smiling, which leads to trustworthiness impressions) override trait-related static facial information (e.g., untrustworthy-looking face). When a less female-stereotypic face, for example, presents a female-stereotypic facial gesture (e.g., smiling), the negative effects of less impression differentiation in female impressions might diminish.

For social justice, our findings suggest another contributing factor to gender discrimination. Women with counterstereotypical appearance are perceived more unfavorably and discriminated more harshly than men with counterstereotypical appearance (Rudman, 1998; Rudman & Phelan, 2008; Sutherland et al., 2015). Given the importance of first impressions (Ballew & Todorov, 2007; Blair et al., 2004; Eberhardt et al., 2006; Funk & Todorov, 2013; Olivola & Todorov, 2010; Todorov et al., 2005), less differentiated impressions of women would result in evaluative inferences that may penalize women more strongly than men for not fitting the expected stereotypes.

In sum, our findings show that people have less differentiated and more valence-laden impressions of women than of men, although these impressions are based on similar visual information. These findings suggest that discrimination against women starts from the moment of forming first impressions, as women with counterstereotypical looks are likely to be evaluated negatively.

References

Abele, A. E. (2003). The dynamics of masculine-agentic and feminine-communal traits: Findings from a prospective study. *Journal of Personality and Social Psychology, 85,* 768–776. http://dx.doi.org/10.1037/0022-3514.85.4.768

Antonakis, J., & Dalgas, O. (2009). *Predicting elections: Child's play! Science, 323,* 1183. http://dx.doi.org/10.1126/science.1167748

Ballew, C. C., II, & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 17948–17953. http://dx.doi.org/10.1073/pnas.0705435104

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

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science, 15,* 674–679. http://dx.doi.org/10.1111/j.0956-7976.2004.00739.x

Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F. E., & Rosenkrantz, P. S. (1972). Sex-role stereotypes: A current appraisal. *Journal of Social Issues, 28,* 59–78. http://dx.doi.org/10.1111/j.1540-4560.1972.tb00018.x

Burriss, R. P., Little, A. C., & Nelson, E. C. (2007). 2D:4D and sexually dimorphic facial characteristics. *Archives of Sexual Behavior, 36,* 377–384. http://dx.doi.org/10.1007/s10508-006-9136-1

Cellerino, A., Borghetti, D., Valenzano, D. R., Tartarelli, G., Mennucci, A., Murri, L., & Sartucci, F. (2007). Neurophysiological correlates for the perception of facial sexual dimorphism. *Brain Research Bulletin, 71,* 515–522. http://dx.doi.org/10.1016/j.brainresbull.2006.11.007

Cislak, A., & Wojciszke, B. (2008). Agency and communion are inferred from actions serving interests of self or others. *European Journal of Social Psychology, 38,* 1103–1110. http://dx.doi.org/10.1002/ejsp.554

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012). Dorsomedial prefrontal cortex mediates rapid evaluations predicting the outcome of romantic interactions. *The Journal of Neuroscience, 32,* 15647–15656. http://dx.doi.org/10.1523/JNEUROSCI.2558-12.2012

Costrich, N., Feinstein, J., Kidder, L., Marecek, J., & Pascale, L. (1975). When stereotypes hurt: Three studies of penalties for sex-role reversals. *Journal of Experimental Social Psychology, 11,* 520–530. http://dx.doi.org/10.1016/0022-1031(75)90003-7

Cundiff, J. L., & Vescio, T. K. (2016). Gender stereotypes influence how people explain gender disparities in the workplace. *Sex Roles: A Journal of Research, 75,* 126–138. http://dx.doi.org/10.1007/s11199-016-0593-2

Deaux, K., & Lewis, L. L. (1984). Structure of gender stereotypes: Interrelationships among components and gender label. *Journal of Personality and Social Psychology, 46,* 991–1004. http://dx.doi.org/10.1037/0022-3514.46.5.991

DeBruine, L. M., & Jones, B. C. (2017). *Face research lab London set*. Retrieved from http://dx.doi.org/10.6084/m9.figshare.5047666

Derlega, V. J., & Chaikin, A. L. (1976). Norms affecting self-disclosure in men and women. *Journal of Consulting and Clinical Psychology, 44,* 376–380. http://dx.doi.org/10.1037/0022-006X.44.3.376

Dotsch, R., Hassin, R. R., & Todorov, A. T. (2016). Statistical learning shapes face evaluation. *Nature Human Behaviour, 1,* 0001. http://dx.doi.org/10.1038/s41562-016-0001

Dotsch, R., & Todorov, A. T. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science, 3,* 562–571. http://dx.doi.org/10.1177/1948550611430272

Eagly, A. H., & Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin, 15,* 543–558. http://dx.doi.org/10.1177/0146167289154008

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

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science, 17,* 383–386. http://dx.doi.org/10.1111/j.1467-9280.2006.01716.x

Ellis, H., Shepherd, J., & Bruce, A. (1973). The effects of age and sex upon adolescents' recognition of faces. *The Journal of Genetic Psychology: Research and Theory on Human Development, 123,* 173–174. http://dx.doi.org/10.1080/00221325.1973.10533202

Engell, A. D., Todorov, A., & Haxby, J. V. (2010). Common neural mechanisms for the evaluation of facial trustworthiness and emotional expressions as revealed by behavioral adaptation. *Perception, 39,* 931–941. http://dx.doi.org/10.1068/p6633

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82,* 878–902. http://dx.doi.org/10.1037/0022-3514.82.6.878

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23,* 1–74. http://dx.doi.org/10.1016/S0065-2601(08)60317-2

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). Looking the part: Social status cues shape race perception. *PLoS ONE, 6*(9), e25107. http://dx.doi.org/10.1371/journal.pone.0025107

Funk, F., & Todorov, A. T. (2013). Criminal stereotypes in the courtroom: Facial tattoos affect guilt and punishment differently. *Psychology, Public Policy, and Law, 19,* 466–478. http://dx.doi.org/10.1037/a0034736

Funk, F., Walker, M., & Todorov, A. T. (2017). Modelling perceptions of criminality and remorse from faces using a data-driven computational approach. *Cognition and Emotion, 31,* 1431–1443. http://dx.doi.org/10.1080/02699931.2016.1227305

Garcia-Retamero, R., & López-Zafra, E. (2006). Prejudice against women in male-congenial environments: Perceptions of gender role congruity in leadership. *Sex Roles, 55,* 51–61. http://dx.doi.org/10.1007/s11199-006-9068-1

Gill, D., Garrod, O. G. B., Jack, R. E., & Schyns, P. G. (2014). Facial movements strategically camouflage involuntary social signals of face morphology. *Psychological Science, 25,* 1079–1086. http://dx.doi.org/10.1177/0956797614522274

Glick, P., & Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70,* 491–512. http://dx.doi.org/10.1037/0022-3514.70.3.491

Glick, P., & Fiske, S. T. (2011). Ambivalent sexism revisited. *Psychology of Women Quarterly, 35,* 530–535. http://dx.doi.org/10.1177/0361684311414832

Glick, P., Fiske, S. T., Mladinic, A., Saiz, J. L., Abrams, D., Masser, B., . . . López, W. (2000). Beyond prejudice as simple antipathy: Hostile and benevolent sexism across cultures. *Journal of Personality and Social Psychology, 79,* 763–775. http://dx.doi.org/10.1037/0022-3514.79.5.763

Goldberg, P. (1968). Are women prejudiced against women? *Trans-Action, 5,* 28–30.

Hagen, R. L., & Kahn, A. (1975). Discrimination against competent women. *Journal of Applied Social Psychology, 5,* 362–376. http://dx.doi.org/10.1111/j.1559-1816.1975.tb00688.x

Hareli, S., Shomrat, N., & Hess, U. (2009). Emotional versus neutral expressions and perceptions of social dominance and submissiveness. *Emotion, 9,* 378–384. http://dx.doi.org/10.1037/a0015958

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10,* 252–264. http://dx.doi.org/10.1207/s15327957pspr1003_4

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,* 513–529. http://dx.doi.org/10.1037/pspa0000090

Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues, 57,* 657–674. http://dx.doi.org/10.1111/0022-4537.00234

Heilman, M. E., & Wallen, A. S. (2010). Wimpy and undeserving of respect: Penalties for men's gender-inconsistent success. *Journal of Experimental Social Psychology, 46,* 664–667. http://dx.doi.org/10.1016/j.jesp.2010.01.008

Hess, U., Blairy, S., & Kleck, R. E. (2000). The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal Behavior, 24,* 265–283. http://dx.doi.org/10.1023/A:1006623213355

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). New York, NY: Guilford Press.

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science, 15,* 342–345. http://dx.doi.org/10.1111/j.0956-7976.2004.00680.x

Imhoff, R., & Koch, A. (2017). How orthogonal are the Big Two of social perception? On the curvilinear relation between agency and communion. *Perspectives on Psychological Science, 12,* 122–137. http://dx.doi.org/10.1177/1745691616657334

Jablonski, N. G., & Chaplin, G. (2000). The evolution of human skin coloration. *Journal of Human Evolution, 39,* 57–106. http://dx.doi.org/10.1006/jhev.2000.0403

Jack, R. E., & Schyns, P. G. (2017). Toward a social psychophysics of face communication. *Annual Review of Psychology, 68,* 269–297. http://dx.doi.org/10.1146/annurev-psych-010416-044242

Jennrich, R. I. (1970). An asymptotic χ^2 test for the equality of two correlation matrices. *Journal of the American Statistical Association, 65,* 904–912.

Keating, C. F., Mazur, A., & Segall, M. H. (1981). A cross-cultural exploration of physiognomic traits of dominance and happiness. *Ethology and Sociobiology, 2,* 41–48. http://dx.doi.org/10.1016/0162-3095(81)90021-2

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative-progressive beliefs, and communion. *Journal of Personality and Social Psychology, 110,* 675–709. http://dx.doi.org/10.1037/pspa0000046

Kramer, R. S. S., Young, A. W., Day, M. G., & Burton, A. M. (2017). Robust social categorization emerges from learning the identities of very few faces. *Psychological Review, 124,* 115–129. http://dx.doi.org/10.1037/rev0000048

Lenz, G. S., & Lawson, C. (2011). Looking the part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science, 55,* 574–589. http://dx.doi.org/10.1111/j.1540-5907.2011.00511.x

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology, 72,* 275–287. http://dx.doi.org/10.1037/0022-3514.72.2.275

Lewin, C., & Herlitz, A. (2002). Sex differences in face recognition—women's faces make the difference. *Brain and Cognition, 50,* 121–128. http://dx.doi.org/10.1016/S0278-2626(02)00016-7

Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior, 28,* 18–27. http://dx.doi.org/10.1016/j.evolhumbehav.2006.09.002

Lundqvist, D., Flykt, A., & Arne, Ö. (1998). *The Karolinska directed emotional faces*. Stockholm, Sweden: Karolinska Hospital.

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,* 1122–1135. http://dx.doi.org/10.3758/s13428-014-0532-5

Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perception and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior, 27,* 237–254. http://dx.doi.org/10.1023/A:1027332800296

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 16474–16479. http://dx.doi.org/10.1073/pnas.1211286109

Moss-Racusin, C. A., Phelan, J. E., & Rudman, L. A. (2010). When men break the gender rules: Status incongruity and backlash against modest men. *Psychology of Men & Masculinity, 11,* 140–151. http://dx.doi.org/10.1037/a0018093

Mouchetant-Rostaing, Y., & Giard, M. H. (2003). Electrophysiological correlates of age and gender perception on human faces. *Journal of Cognitive Neuroscience, 15*(Suppl.), 900–910. http://dx.doi.org/10.1162/089892903322370816

Mouchetant-Rostaing, Y., Giard, M. H., Bentin, S., Aguera, P. E., & Pernier, J. (2000). Neurophysiological correlates of face gender processing in humans. *The European Journal of Neuroscience, 12,* 303–310. http://dx.doi.org/10.1046/j.1460-9568.2000.00888.x

Nussbaum, M. C. (1999). *Sex and social justice*. Oxford, UK: Oxford University Press.

Oh, D., Buck, E. A., & Todorov, A. (2019). Revealing hidden gender biases in competence impressions from faces. *Psychological Science, 30,* 65–79. http://dx.doi.org/10.1177/0956797618813092

Oh, D., Shafir, E., & Todorov, A. (2019). *Economic status cues from clothes affect perceived competence from faces*. Retrieved from http://dx.doi.org/10.31234/osf.io/saqnv

Olivola, C. Y., & Todorov, A. T. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior, 34,* 83–110. http://dx.doi.org/10.1007/s10919-009-0082-1

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

Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion, 9,* 128–133. http://dx.doi.org/10.1037/a0014520

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Parks-Stamm, E. J., Heilman, M. E., & Hearns, K. A. (2008). Motivated to penalize: Women's strategic rejection of successful women. *Personality and Social Psychology Bulletin, 34,* 237–247. http://dx.doi.org/10.1177/0146167207310027

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., . . . Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature, 394,* 884–887. http://dx.doi.org/10.1038/29772

Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly, 26,* 269–281. http://dx.doi.org/10.1111/1471-6402.t01-1-00066

Prentice, D. A., & Carranza, E. (2004). Sustaining cultural beliefs in the face of their violation: The case of gender stereotypes. In M. Schaller & C. S. Crandall (Eds.), *The psychological foundations of culture* (pp. 259–280). Mahwah, NJ: Erlbaum.

Rehnman, J., & Herlitz, A. (2006). Higher face recognition ability in girls: Magnified by own-sex and own-ethnicity bias. *Memory, 14,* 289–296. http://dx.doi.org/10.1080/09658210500233581

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

Rhodes, G., Hickford, C., & Jeffery, L. (2000). Sex-typicality and attractiveness: Are supermale and superfemale faces super-attractive? *British Journal of Psychology, 91,* 125–140. http://dx.doi.org/10.1348/000712600161718

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9,* 283–294. http://dx.doi.org/10.1037/h0026086

Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology, 74,* 629–645. http://dx.doi.org/10.1037/0022-3514.74.3.629

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57,* 743–762. http://dx.doi.org/10.1111/0022-4537.00239

Rudman, L. A., & Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in Organizational Behavior, 28,* 61–79. http://dx.doi.org/10.1016/j.riob.2008.04.003

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

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

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features! Understanding recognition from the use of visual information. *Psychological Science, 13,* 402–409. http://dx.doi.org/10.1111/1467-9280.00472

Secord, P. F., & Berscheid, E. S. (1963). Stereotyping and the generality of implicit personality theory. *Journal of Personality, 31,* 65–78. http://dx.doi.org/10.1111/j.1467-6494.1963.tb01841.x

Sofer, C., Dotsch, R., Wigboldus, D. H. J., & Todorov, A. (2015). What is typical is good: The influence of face typicality on perceived trustworthiness. *Psychological Science, 26,* 39–47. http://dx.doi.org/10.1177/0956797614554955

South Palomares, J. K., Sutherland, C. A. M., & Young, A. W. (2018). Facial first impressions and partner preference models: Comparable or distinct underlying structures? *British Journal of Psychology, 109,* 538–563.

Spence, J. T., Helmreich, R. L., & Holahan, C. K. (1979). Negative and positive components of psychological masculinity and femininity and their relationships to self-reports of neurotic and acting out behaviors. *Journal of Personality and Social Psychology, 37,* 1673–1682. http://dx.doi.org/10.1037/0022-3514.37.10.1673

Spence, J. T., Helmreich, R., & Stapp, J. (1975). Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology, 32,* 29–39. http://dx.doi.org/10.1037/h0076857

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience, 19,* 795–797. http://dx.doi.org/10.1038/nn.4296

Stolier, R. M., & Freeman, J. B. (2017). A neural mechanism of social categorization. *The Journal of Neuroscience, 37,* 5711–5721. http://dx.doi.org/10.1523/JNEUROSCI.3334-16.2017

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A dynamic structure of social trait space. *Trends in Cognitive Sciences, 22,* 197–200. http://dx.doi.org/10.1016/j.tics.2017.12.003

Sutherland, C. A. M., Liu, X., Zhang, L., Chu, Y., Oldmeadow, J. A., & Young, A. W. (2018). Facial first impressions across culture: Data-driven modeling of Chinese and British perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin, 44,* 521–537.

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127,* 105–118. http://dx.doi.org/10.1016/j.cognition.2012.12.001

Sutherland, C. A. M., Oldmeadow, J. A., & Young, A. W. (2016). Integrating social and facial models of person perception: Converging and diverging dimensions. *Cognition, 157,* 257–267. http://dx.doi.org/10.1016/j.cognition.2016.09.006

Sutherland, C. A. M., Rhodes, G., & Young, A. W. (2017). Facial image manipulation: A tool for investigating social perception. *Social Psychological and Personality Science, 8,* 538–551. http://dx.doi.org/10.1177/1948550617697176

Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (2015). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology, 106,* 186–208. http://dx.doi.org/10.1111/bjop.12085

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

Tiddeman, B., Burt, M., & Perrett, D. I. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications, 21,* 42–50. http://dx.doi.org/10.1109/38.946630

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

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion, 13,* 724–738. http://dx.doi.org/10.1037/a0032335

Todorov, A. T., Dotsch, R., Wigboldus, D. H. J., & Said, C. P. (2011). Data-driven methods for modeling social perception. *Social and Personality Psychology Compass, 5,* 775–791. http://dx.doi.org/10.1111/j.1751-9004.2011.00389.x

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308,* 1623–1626. http://dx.doi.org/10.1126/science.1110589

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66,* 519–545. http://dx.doi.org/10.1146/annurev-psych-113011-143831

Todorov, A. T., & Oosterhof, N. N. (2011). Modeling social perception of faces. *IEEE Signal Processing Magazine, 28,* 117–122. http://dx.doi.org/10.1109/MSP.2010.940006

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12,* 455–460. http://dx.doi.org/10.1016/j.tics.2008.10.001

Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision, 9*(11), 12. http://dx.doi.org/10.1167/9.11.12

Walker, M., & Vetter, T. (2016). Changing the personality of a face: Perceived Big Two and Big Five personality factors modeled in real photographs. *Journal of Personality and Social Psychology, 110,* 609–624. http://dx.doi.org/10.1037/pspp0000064

Welling, L. L. M., Bestelmeyer, P. E. G., Jones, B. C., DeBruine, L. M., & Allan, K. (2017). Effects of sexually dimorphic shape cues on neurophysiological correlates of women's face processing. *Adaptive Human Behavior and Physiology, 3,* 337–350. http://dx.doi.org/10.1007/s40750-017-0072-1

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology, 37,* 395–412. http://dx.doi.org/10.1037/0022-3514.37.3.395

Williams, J. E., & Best, D. L. (1990). *Sex and psyche: Gender and self viewed cross-culturally*. Thousand Oaks, CA: Sage.

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science, 26,* 1325–1331. http://dx.doi.org/10.1177/0956797615590992

Zebrowitz, L. A. (2004). The origin of first impressions. *Journal of Cultural and Evolutionary Psychology, 2,* 93–108. http://dx.doi.org/10.1556/JCEP.2.2004.1-2.6

Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior, 15,* 603–623. http://dx.doi.org/10.1007/BF01065855

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass, 2,* 1497–1517. http://dx.doi.org/10.1111/j.1751-9004.2008.00109.x

Received April 12, 2018
Revision received May 13, 2019
Accepted May 17, 2019 ■