Research Article

Psychological Science
2019, Vol. 30(1) 65–79

DOI: 10.1177/0956797618813092
www.psychologicalscience.org/PS

Revealing Hidden Gender Biases in Competence Impressions of Faces

DongWon Oh, Elinor A. Buck, and Alexander Todorov
Department of Psychology, Princeton University

Abstract

Competence impressions from faces affect important decisions, such as hiring and voting. Here, using data-driven computational models, we identified the components of the competence stereotype. Faces manipulated by a competence model varied in attractiveness (Experiment 1a). However, faces could be manipulated on perceived competence controlling for attractiveness (Experiment 1b); moreover, faces perceived as more competent but not attractive were also perceived as more confident and masculine, suggesting a bias to perceive male faces as more competent than female faces (Experiment 2). Correspondingly, faces manipulated to appear competent but not attractive were more likely to be classified as male (Experiment 3). When masculinity cues that induced competence impressions were applied to real-life images, these cues were more effective on male faces (Experiment 4). These findings suggest that the main components of competence impressions are attractiveness, confidence, and masculinity, and they reveal gender biases in how we form important impressions of other people.

Keywords

face perception, facial features, stereotypes, gender, open data, open materials

Received 1/28/18; Revision accepted 8/15/18

First impressions from facial appearance are formed effortlessy and shape significant social outcomes (Todorov, 2017; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). Impressions of competence are especially important, because they influence decisions about leadership selection (Antonakis & Eubanks, 2017). Intuitive judgments of competence from faces, for instance, can predict the results of political elections (Antonakis & Dalgas, 2009; Ballew & Todorov, 2007; Lenz & Lawson, 2011; Olivola & Todorov, 2010; Todorov, Mandisodza, Goren, & Hall, 2005) and company executives' compensation (Graham, Harvey, & Puri, 2017; Stoker, Garretsen, & Spreeuwers, 2016). It is important to understand the perceptual basis of these impressions, because people act on these impressions (e.g., choose their leaders on the basis of competence impressions) despite the dubious relationship between leaders' actual competence and competence impressions from their faces (Stoker et al., 2016; Wyatt & Silvester, 2018).

Here, we investigated the visual ingredients of the competence stereotype. Facial attractiveness is one of these ingredients. Both empirical studies and computational models of facial impressions support the "halo effect" of attractiveness (Dion, Berscheid, & Walster, 1972; Landy & Sigall, 1974; Thorndike, 1920) on competence impressions. First, a meta-analysis showed a modest to strong association between attractiveness and perceived social and intellectual competence (Eagly, Ashmore, Makhijani, & Longo, 1991). Individuals with attractive faces are perceived as socially and occupationally competent (Dion et al., 1972; Landy & Sigall, 1974) and as having a higher social status (Webster & Driskell, 2015), which is strongly associated with perceived competence (Fiske, Cuddy, Glick, & Xu, 2002). In real-world data, judgments of competence and attractiveness from politicians' faces (N = 244) are highly correlated (Olivola & Todorov, 2010). Second, data-driven models of facial impressions (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011) show a strong similarity between models of competence and attractiveness (Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013). Because the two models exist in a common space, one can directly assess the similarities between

Corresponding Author:
DongWon Oh, Princeton University, Department of Psychology, 322 Peretsman-Scully Hall, Princeton, NJ 08544
E-mail: dongwoh@gmail.com

them: The models of competence and attractiveness are indeed highly similar ($\rho = .71$), suggesting that people rely on attractiveness when forming impressions of competence.

We tested whether there are meaningful visual components other than attractiveness that contribute to competence impressions. Data-driven computational models of impressions (Todorov et al., 2013; Todorov & Oosterhof, 2011) are particularly suitable for addressing this question. Because the competence and attractiveness models are in the same statistical space, we can create a new competence model that is not confounded with attractiveness by either (a) making the new competence model statistically orthogonal (uncorrelated) to the attractiveness model or (b) forcing the new model to be negatively correlated with the attractiveness model by subtracting the attractiveness model from the competence model. To the extent that the new competence model (e.g., the resulting competence-minus-attractivness model) is meaningful, faces that are perceived as more competent should not be perceived as more attractive. More importantly, if the model still predicts competence impressions, then by inspecting this model, we can find out meaningful components of competence impressions that are not readily apparent.

One potential component of these impressions is facial masculinity. When asked to evaluate one's self and others on multiple attributes, people evaluate men as more competent (Bem, 1974; Broverman, Vogel, Broverman, Clarkson, & Rosenkrantz, 1972; Spence, Helmreich, & Stapp, 1975) and more confident (Broverman et al., 1972; Spence et al., 1975) than women, on average. Further, the beliefs in the association between men and competence, confidence, and semantically similar traits (e.g., independence, inventiveness) are held across diverse cultures (Williams & Best, 1990). However, the influence of masculinity on competence impressions may not be immediately apparent in the model of competence, because attractiveness is highly positively correlated with feminine facial appearance in both genders (Perrett et al., 1998; Rhodes, Hickford, & Jeffery, 2000; Said & Todorov, 2011; but see Rhodes, 2006). By controlling for the attractiveness of faces, we can directly test whether masculinity contributes to competence impressions.

Following this logic, we tested for gender biases in competence impressions and uncovered multiple components underlying these impressions. In Experiment 1a, we showed that both judgments of attractiveness and competence change as faces are manipulated to look more competent by the standard competence model (Todorov et al., 2013). In Experiment 1b, we created a new model of competence by subtracting the model of attractiveness and showed that faces manipulated by this model to look competent are indeed perceived to be more competent but not more attractive. More importantly, we showed in Experiment 2 that these faces are also perceived as more masculine and confident. In Experiment 3, we showed that the competent-looking faces are more likely to be categorized as men than as women and that the incompetent-looking faces are more likely to be categorized as women. In Experiment 4, we extended our findings to real-life face images. We showed that whereas masculinity cues increase competence impressions of male faces, they increase competence impressions of female faces only up to a point, after which they decrease their perceived competence.

Experiment 1a

Using a validated, data-driven, computational model of competence (Todorov et al., 2013), we manipulated faces to vary in their perceived competence (see Fig. 1). Participants were asked to evaluate these faces either on competence or on attractiveness. Given the high positive correlation between attractiveness impressions and competence impressions (e.g., Dion et al., 1972; Olivola & Todorov, 2010; Todorov et al., 2013), we expected that as the level of model manipulation increased, both competence and attractiveness ratings of the faces would increase.

Method

Participants. Thirty-three participants (17 men, 16 women; age: $M = 35.30$ years, range = 24–58) were recruited via Amazon Mechanical Turk (MTurk) and participated for payment. A power analysis using G*Power (Version 3.1.9; Faul, Erdfelder, Lang, & Buchner, 2007) indicated that a sample size of at least 10 participants per condition would afford 95% power to detect a large effect ($R^2 > .65$) of impression-manipulation levels in participant-level regressions, which was found in a previous validation study with the same design (Todorov et al., 2013). In the previous study, the manipulation level of impression models, including that of the competence and attractiveness models, predicted a systematic change in impression ratings of the faces in linear and quadratic regression models.

Materials. We generated face stimuli with FaceGen (Singular Inversions, Toronto, Canada) from a data-driven model of impressions of competence (Todorov & Oosterhof, 2011). A data-driven model extracts the visual information used to form an impression without constraining the search to a priori set of facial features (Jack & Schyns, 2017). In the FaceGen model, each face is a vector with 100 parameters in a 100-dimensional face space. A change in a parameter

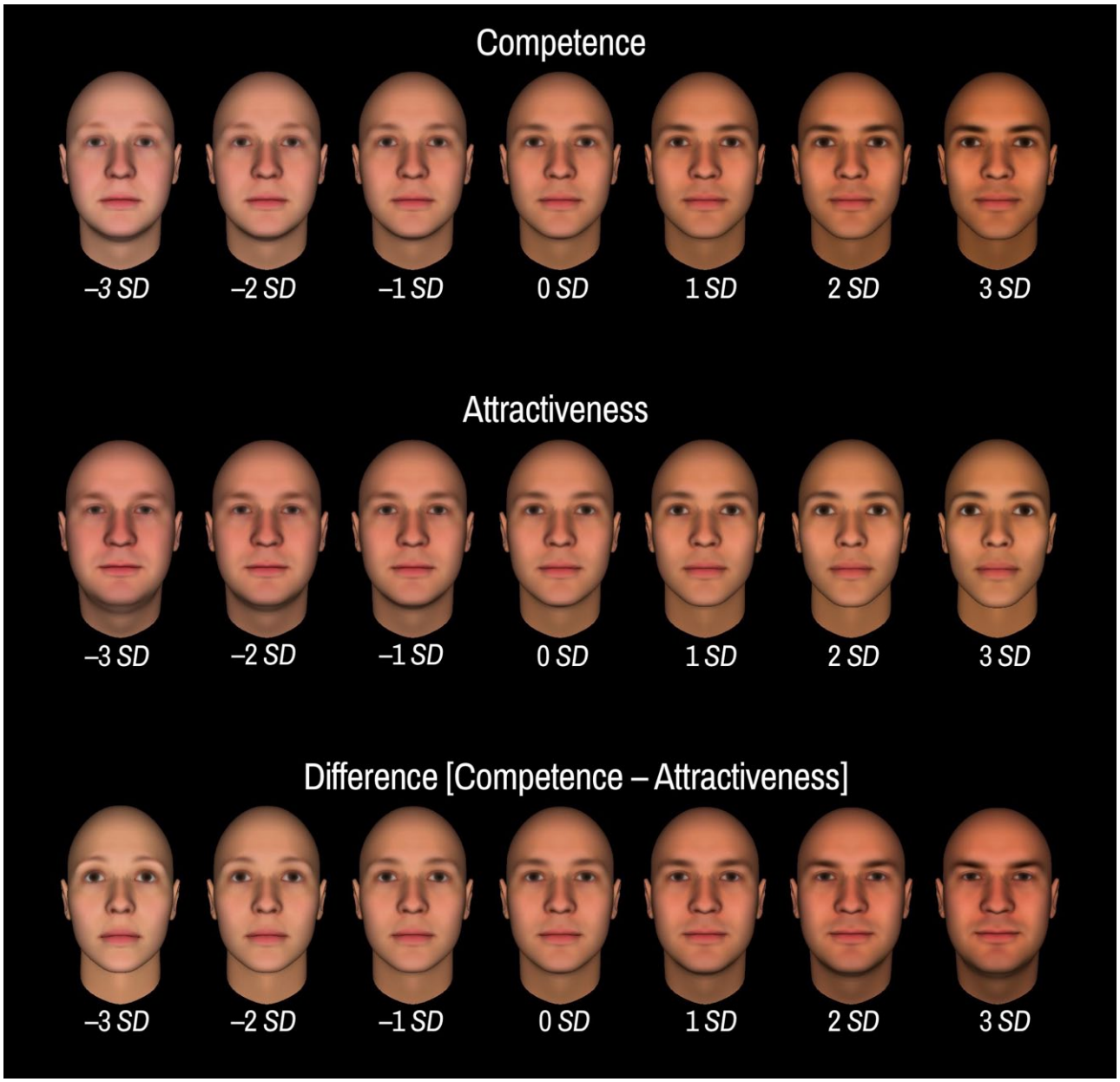

Fig. 1. A face manipulated by the competence model (top), the attractiveness model (middle), and the difference model (competence – attractiveness; bottom). As the standard-deviation units increase, the face is perceived as more competent (top), more attractive (middle), and more competent but not attractive (bottom).

causes a holistic change in facial appearance, which is orthogonal to changes caused by other parameters. Fifty of the 100 parameters determine the shape of a face, and the 50 other parameters determine its reflectance (i.e., texture and pigmentation). One can calculate the contribution of each parameter to this impression on the basis of people's ratings of many ($N \geq 300$) randomly generated faces on a single social impression (e.g., competence), and consequently, one can model this impression as a linear vector in the statistical face space (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011). With this method, Todorov et al. (2013) created and validated a model of competence impressions and a model of attractiveness impressions, among other models. In this research, we used these two models. In Experiment 1a, we used the standard competence model (Todorov et al.,

2013). For Experiments 1b, 2, 3, and 4, we used a new model created by subtracting the attractiveness model from the competence model (we also used an orthogonal model in four additional experiments; see the Supplemental Material available online). Figure 1 demonstrates how the average FaceGen face changes when it is manipulated by the models of competence, attractiveness, and their difference.

Before applying the computational models to faces, we created 25 new, distinct faces (see Fig. S1 in the Supplemental Material). These 25 faces were chosen from a random sample of 1,000 faces as those with the largest differences, on the basis of the average euclidean distance between the faces (Todorov et al., 2013). Because the process of choosing distinct faces resulted in 25 atypical faces, the faces were scaled to more closely resemble the average face while still preserving the ratio of differences between them. Then, for each identity, we generated 7 faces: Each identity was projected at −3, −2, −1, 0, 1, 2, and 3 standard deviations on the dimension of the competence model. Each standard-deviation value represents the expected amount of change in competence ratings that would be caused by the corresponding change in the appearance of the face relative to the average face (Todorov et al., 2013). The complete set of stimuli consisted of 175 face images (25 identities × 7 manipulation levels).

Procedure. Each participant was randomly assigned to make judgments of either competence or attractiveness. The stimuli were presented in random order to each participant. For each stimulus, the question asked was "How [trait] is this person?" which participants responded to using a 9-point scale ranging from 1 (*not at all [trait]*) to 9 (*extremely [trait]*). The participants were shown only one face for each trial and were blind to the impression model by which the faces had been manipulated. Before the study began, each participant was told to rely on "gut instinct," not to spend too much time on each face, and that there were no right or wrong answers. Participants were given unlimited time for each trial.

To assess intrarater reliability, we added 25 repeated trials randomly chosen from the first 175 trials in each study. These extra 25 trials brought the total number of trials to 200. Using the ratings from the 25 repeated trial pairs, we calculated a correlational coefficient as a measure of test-retest reliability of each participant. We excluded from further analyses the responses of any participants with test-retest reliability less than or equal to 0: 1 participant in the competence-rating condition and 2 participants in the attractiveness-rating condition. We recruited additional participants so that we had 15 participants with test-retest reliability greater than 0 per impression. The interrater reliabilities of the impression ratings were high (competence: $\alpha = .90$, attractiveness: $\alpha = .90$).

Results

To test whether competence and attractiveness impressions tracked the competence model manipulation, we fitted linear and quadratic regression models for the impression ratings. For the regressions, the impression ratings were averaged across participants (face-level analysis, $n = 25$ each impression) and across face identities (participant-level analysis, $n = 15$ each impression). The fit was good across all models, showing that the judgments were well explained as a function of the manipulation level (see Fig. 2).

The effect of the model manipulation on the competence and attractiveness ratings was consistent across face identities (see Figs. 2 and S2 in the Supplemental Material). The linear model explained more than 75% of the variance in the ratings—competence: $R^2 = .83$, $F(1, 173) = 824.30$, $p < .001$; attractiveness: $R^2 = .79$, $F(1, 173) = 661.82$, $p < .001$. Although both competence and attractiveness ratings were explained by the competence manipulation, when we compared the coefficients from the regression models for competence and attractiveness ratings using multivariate models (Zellner, 1963), the face-impression manipulation induced a bigger change in the competence ratings ($b_1 = 0.56$) than in the attractiveness ratings ($b_1 = 0.50$; $z = 3.20$, $p = .001$). This is not surprising given that the competence model had been built to capture the facial information underlying perceived competence, not attractiveness.

The quadratic model explained more than 80% of the variance—competence: $R^2 = .86$, $F(2, 172) = 517.01$, $p < .001$; attractiveness: $R^2 = .80$, $F(2, 172) = 354.06$, $p < .001$. The quadratic fits were better than the linear fits—competence: $F(1, 172) = 37.21$, $p < .001$; attractiveness: $F(1, 172) = 10.39$, $p = .002$. When we compared the coefficients from the regression models for competence and attractiveness ratings, the face-impressions manipulation again induced a bigger change in the competence ratings than in the attractiveness ratings for both the quadratic terms (competence: $b_2 = -0.06$, attractiveness: $b_2 = -0.04$; $z = 3.11$, $p = .002$) and linear terms (competence: $b_1 = 1.06$, attractiveness: $b_1 = 0.78$; $z = 2.48$, $p = .013$).

The results were similar when the analysis was conducted at the level of participants (see Figs. S3 and S4 in the Supplemental Material). The linear models explained more than 50% of the variance in the ratings—competence: $R^2 = .54$, $F(1, 103) = 122.29$, $p < .001$; attractiveness: $R^2 = .57$, $F(1, 103) = 135.88$, $p < .001$. The quadratic models explained more than 55% of the variance—competence: $R^2 = .56$, $F(2, 102) =$

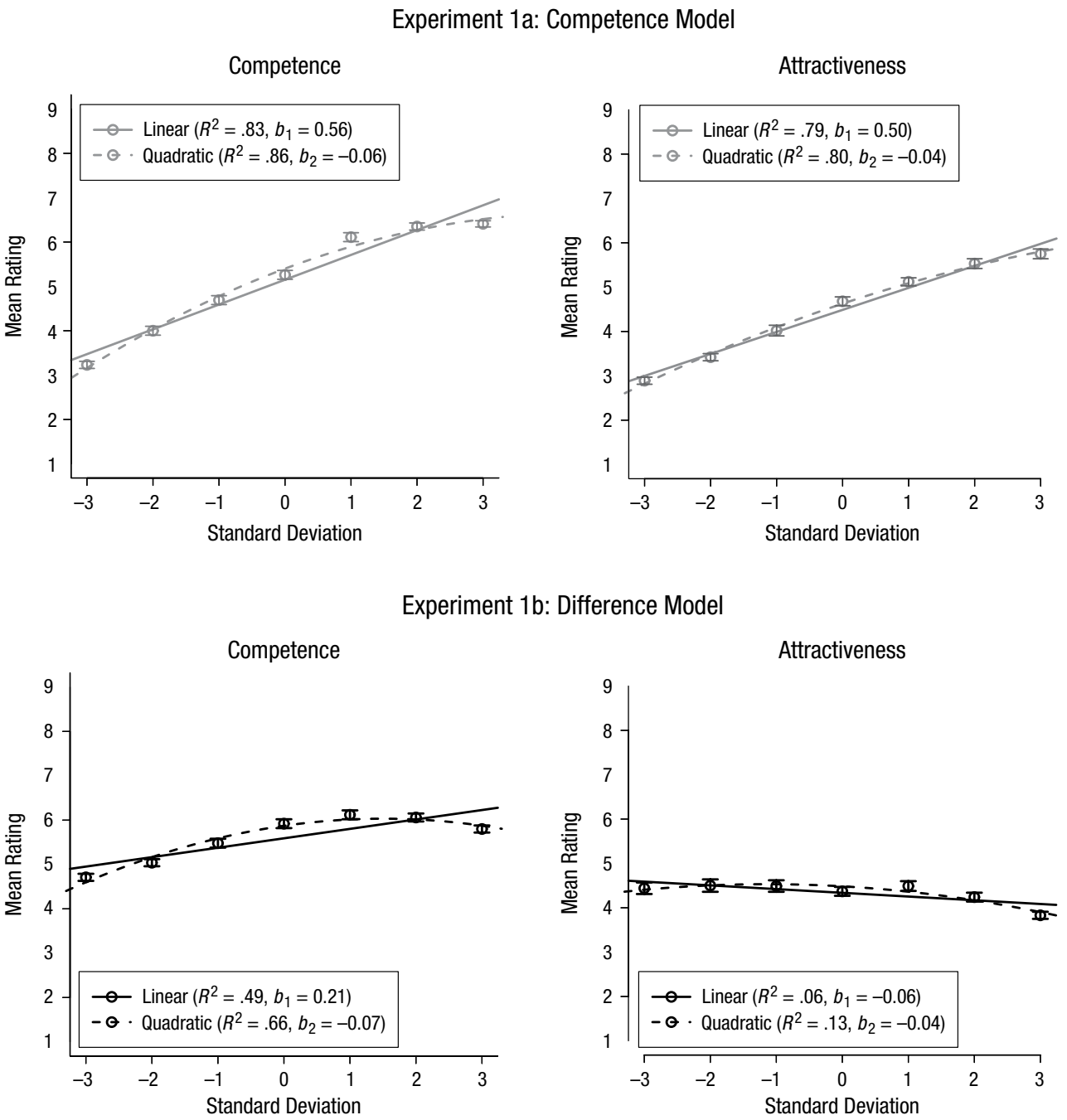

Fig. 2. Results of Experiment 1a (top) and Experiment 1b (bottom). The mean impression ratings of competence (left) and attractiveness (right) are shown as a function of the level of the competence-model manipulation (Experiment 1a) and the difference-model manipulation (Experiment 1b). For each impression, fits of the linear and quadratic models are shown for the mean rating, averaged across participants. Error bars denote standard errors.

65.75, $p < .001$; attractiveness: $R^2 = .58$, $F(2, 102) = 69.66$, $p < .001$.

These results show that when facial cues of competence impressions are enhanced, both competence and attractiveness impressions increase together. This covariance suggests that the visual components of competence and attractiveness impressions from faces are highly confounded with each other under natural circumstances. This confounding relationship is consistent with the idea that the halo effect can explain competence impressions.

Experiment 1b

The results of Experiment 1a show that facial attractiveness is a major ingredient of competence impressions.

However, it is unclear whether there are other meaningful ingredients when attractiveness is not positively correlated with competence impressions. In Experiment 1b, we created a new model—that is, the difference between the competence and attractiveness models (referred to hereafter as the *difference model*)—and applied this model to new faces. Theoretically, this model should force judgments of competence and attractiveness to be negatively correlated. However, the mapping between the model space and the psychological judgment space may not be linear (Oosterhof & Todorov, 2008; we obtained results using a competence model orthogonal to the attractiveness model, too; see Experiment S1 in the Supplemental Material). To test how judgments of competence and attractiveness change as a function of the difference model, we asked participants to evaluate faces manipulated by this model on either competence or attractiveness.

Method

Participants. One hundred twenty-five MTurk workers (72 men, 53 women; age: $M = 37.05$ years, range = 18–70) participated for payment. A power analysis using G*Power 3.1.9 indicated that a sample size of 45 participants per condition would afford 90% power to detect a medium effect ($R^2 = .20$) of impression-manipulation levels in participant-level regressions. We expected a medium effect size because removing attractiveness from the competence model (the resulting difference model) should attenuate the effects of the model manipulation on judgments.

Materials. To create the difference model, we subtracted each of the 100 parameters defining the attractiveness model from each of the 100 parameters defining the competence model. To the extent that the difference model works, faces manipulated to be perceived as more competent should also be perceived as less attractive than faces manipulated to be perceived as less competent, or at the very least as attractive as them.

The same 25 identities created for Experiment 1a were employed (see Fig. S1). Each identity was projected at −3, −2, −1, 0, 1, 2, and 3 standard deviations on the dimension of the difference model, resulting in 7 faces per identity. As a result, as in Experiment 1a, the complete set of stimuli consisted of 175 face images (25 identities × 7 manipulation levels).

Procedure. The procedure was identical to that used in Experiment 1a except that faces were created using the difference model. As we did in Experiment 1a, we excluded from further analyses the responses of any participants with test-retest reliability less than or equal to 0: 18 participants in the competence-rating condition and 17 participants in the attractiveness-rating condition. We recruited additional participants so that we had 45 participants with test-retest reliability greater than 0 per impression condition. The interrater reliabilities were high (competence: $\alpha = .82$, attractiveness: $\alpha = .75$).

Results

Linear and quadratic regression models were fitted for the impression ratings to test whether competence and attractiveness impressions tracked the difference-model manipulation. For the regressions, the impression ratings were averaged across participants (face-level analysis, $n = 25$ each impression) and across face identities (participant-level analysis, $n = 45$ each impression). The model fit was good for the competence judgments but not the attractiveness judgments, showing that only the competence judgments were well explained as a function of the manipulation level (see Fig. 2). As the model-manipulation level increased, ratings of competence increased, too, but ratings of attractiveness did not.

The effect of the model manipulation on the competence and attractiveness ratings was consistent across face identities (see Figs. 2 and S2). However, the impression manipulation was far more impactful on the competence ratings than on the attractiveness ratings. The linear models explained 49% of the variance in the competence ratings, $F(1, 173) = 163.04$, $p < .001$, but only 6% of the variance in the attractiveness ratings, $F(1, 173) = 11.61$, $p < .001$. When we compared the coefficients from the regression models for competence and attractiveness, the manipulation induced a far bigger change in the competence ratings ($b_1 = 0.21$) than in the attractiveness ratings ($b_1 = -0.06$; $z = 17.39$, $p < .001$). This finding shows that the difference model was indeed capable of varying the faces' perceived competence without varying their attractiveness ratings too much. If anything, more competent-looking faces were perceived to be less attractive.

The quadratic model explained 66% of the variance in the competence ratings, $F(2, 172) = 165.52$, $p < .001$, but only 13% of the variance in the attractiveness ratings, $F(2, 172) = 13.26$, $p < .001$. The quadratic fits were better than the linear fits—competence: $F(1, 172) = 86.98$, $p < .001$; attractiveness: $F(1, 172) = 14.03$, $p < .001$. When we compared the coefficients from the regression models for competence and attractiveness ratings, the face-impressions manipulation again induced a bigger change in the competence ratings than in the attractiveness ratings for both the quadratic terms (competence: $b_2 = -0.07$, attractiveness: $b_2 = -0.04$; $z = 4.40$, $p < .001$) and linear terms (competence: $b_1 = 0.79$, attractiveness: $b_1 = 0.23$; $z = 8.17$, $p < .001$).

The results were consistent with the analysis conducted at the level of participants (see Figs. S3 and S4). The linear models explained 8% of the variance in the

competence ratings, $F(1, 313) = 26.27$, $p < .001$, but an insignificant amount (< 1%) of the variance in the attractiveness ratings, $F(1, 313) = 1.36$, $p = .245$. The quadratic models explained 11% of the variance in the competence ratings, $F(2, 312) = 18.32$, $p < .001$, but an insignificant amount (< 1%) of the variance in the attractiveness ratings, $F(2, 312) = 1.45$, $p = .237$.

The results show that when facial cues of perceived competence are enhanced by the difference model, competence impressions increased but attractiveness impressions decreased (face-level analysis) or did not vary at all (participant-level analysis). This negative or null correlation between competence and attractiveness impressions contrasts with the high positive correlation between these impressions when facial cues of perceived competence were manipulated by the standard model (Experiment 1a; we obtained results using a competence model orthogonal to the attractiveness model, too. The orthogonal model could not control for the halo effect of attractiveness; see Experiment S1). These results show that perceived competence can be meaningfully manipulated, controlling for the halo effect of attractiveness.

Experiment 2

Visual inspection of the difference model (see Fig. 1) shows that as the faces increase in perceived competence, but not attractiveness, they express more confidence and look more masculine. This is consistent with prior research showing strong associations between competence impressions, confidence impressions, and gender (e.g., Spence et al., 1975), as well as research showing high correlations between femininity and attractiveness (e.g., Said & Todorov, 2011). To formally test whether facial confidence and masculinity underlie competence impressions, we asked participants to evaluate faces varying on perceived competence but not attractiveness on either masculinity or confidence.

Method

Participants. Ninety-eight MTurk workers (48 men, 49 women, 1 nonbinary/other gender; age: $M = 38.31$ years, range = 21–70) participated for payment. A power analysis using G*Power 3.1.9 indicated that a sample size of 45 participants per condition would afford 90% power to detect a medium effect ($R^2 = .20$) of impression-manipulation levels in participant-level regressions.

Materials. The same 25 identities created for Experiments 1a and 1b were employed (see Fig. S1). Each identity was projected at −3, −2, −1, 0, 1, 2, and 3 standard deviations on the dimension of the difference model, resulting in 7 faces per identity. As a result, as in Experiment 1b, the complete set of stimuli consisted of 175 face images (25 identities × 7 manipulation levels).

Procedure. The procedure was identical to that used in the previous experiments except that each participant was randomly assigned to make judgments of either confidence or masculinity. As in previous experiments, we excluded from further analyses the responses of any participants with test-retest reliability less than or equal to 0: 6 participants in the confidence-rating condition and 2 participants in the masculinity-rating condition. We recruited additional participants so that we had 45 participants with test-retest reliability greater than 0 per impression condition. The interrater reliabilities were high (confidence: $\alpha = .96$, masculinity: $\alpha = .98$).

Results

Linear and quadratic regression models were fitted for the impression ratings to test whether the confidence- and masculinity-impression ratings tracked the difference-model manipulation. For the regressions, the impression ratings were averaged across participants (face-level analysis, $n = 25$ each impression) and across face identities (participant-level analysis, $n = 45$ each impression). The fit was good across all models, showing that the judgments were well explained as a function of the manipulation level (see Fig. 3).

The effect of the model manipulation on the confidence and masculinity ratings was consistent across face identities (see Figs. 3 and S5 in the Supplemental Material). The linear model explained more than 85% of the variance in the ratings—confidence: $R^2 = .88$, $F(1, 173) = 1{,}283.66$, $p < .001$; masculinity: $R^2 = .92$, $F(1, 173) = 2{,}045.29$, $p < .001$. The quadratic model explained more than 85% of the variance—confidence: $R^2 = .90$, $F(2, 172) = 747.8$, $p < .001$; masculinity: $R^2 = .94$, $F(2, 172) = 1{,}336.79$, $p < .001$. The quadratic fits were better than the linear fits—confidence: $F(1, 172) = 26.06$, $p < .001$; masculinity: $F(1, 172) = 49.92$, $p < .001$.

The results were similar when the analysis was conducted at the level of participants (see Figs. S6 and S7 in the Supplemental Material). The linear model explained more than 45% of the variance in the ratings—confidence: $R^2 = .45$, $F(1, 313) = 259.02$, $p < .001$; masculinity: $R^2 = .73$, $F(1, 313) = 829.15$, $p < .001$. The quadratic model also explained more than 45% of the variance—confidence: $R^2 = .46$, $F(2, 312) = 133.33$, $p < .001$; masculinity: $R^2 = .74$, $F(2, 312) = 443.46$, $p < .001$.

The results show that when facial cues of perceived competence are enhanced by the difference model, both confidence and masculinity impressions increase. These relationships were expected from the visual

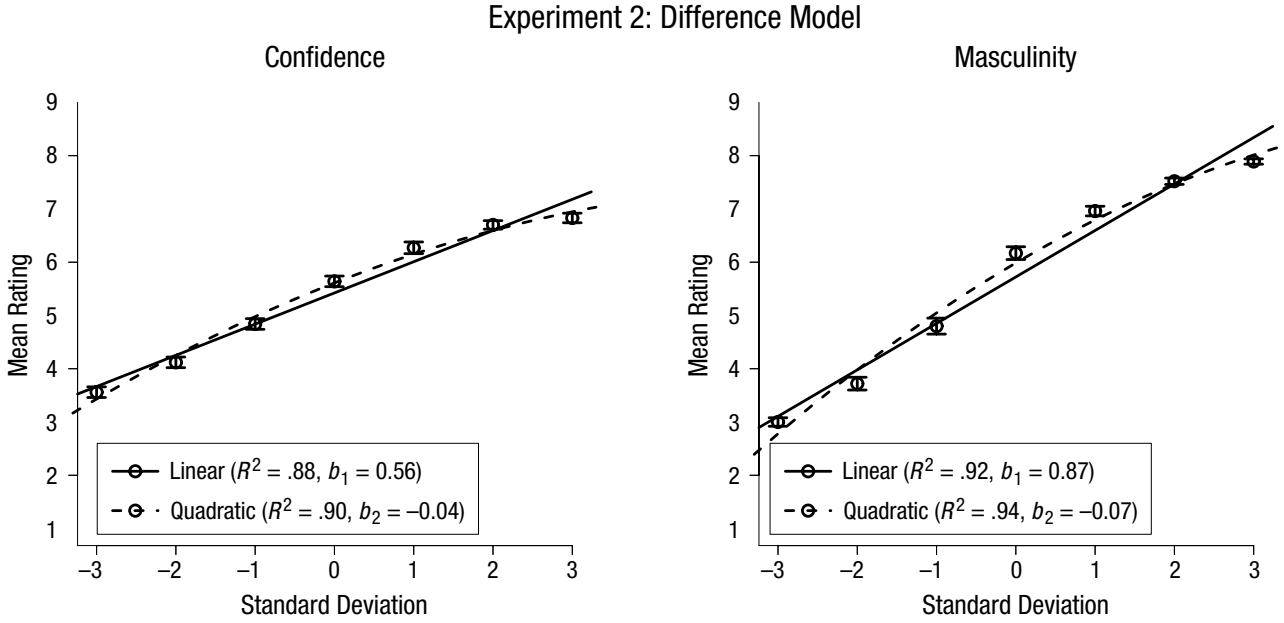

Fig. 3. Results of Experiment 2. The mean impression ratings of confidence (left) and masculinity (right) are shown as a function of the level of the difference-model manipulation. For each impression, fits of the linear and quadratic models are shown for the mean rating, averaged across participants. Error bars denote standard errors.

inspection of the model (see Fig. 1) and the previous literature, as discussed earlier (we obtained similar results using a competence model orthogonal to the attractiveness model; see Experiment S2 in the Supplemental Material). Competence and attractiveness impressions are not positively correlated in the faces used here, as shown in Experiment 1b. It follows that the variance in the competence impressions cannot be attributed to the halo effect of attractiveness, which is a significant natural confound of competence impressions. Thus, the results show that confidence and masculinity cues are important ingredients of competence impressions—ingredients that cannot be explained as a by-product of attractiveness.

Experiment 3

Masculinity cues are strongly related to perceptions of gender, which suggests the presence of gender biases in competence impressions. To directly test whether people use gender-related facial cues to judge competence, we asked participants to categorize faces varying on perceived competence as male or female. We used faces manipulated by both the standard competence model and the difference model. Given the high positive correlation between competence impressions and masculinity, especially in the absence of the halo effect, we expected that (a) participants would be more likely to categorize the faces as male as the level of model manipulation increases, irrespective of the model (i.e., the standard competence or the difference model), and (b) this effect would be accentuated for faces generated by the difference model. Specifically, controlling for attractiveness, we expected that faces manipulated to be perceived as less competent would be more likely to be categorized as female.

Method

Participants. Thirty-one MTurk workers (22 men, 9 women; age: M = 36.32 years, range = 20–58) participated for payment. A power analysis using G*Power 3.1.9 indicated that a sample size of at least 26 participants would afford 95% power to detect a small to medium effect (f = .25) of the main effects of manipulation level and model type, as well as the interaction between the two.

Materials. We used both the face images manipulated by the competence model and the face images manipulated by the difference model. This created a combined pool of 350 face-image stimuli (2 models × 25 identities × 7 manipulation levels).

Procedure. Participants were asked to make a forced choice of perceived gender for each face. All participants were exposed to faces from both the competence and difference models. Two versions of the study with the same length were created: Half of the participants were presented with 88 competence-model faces and 87 difference-model faces, whereas the other half were presented with

87 competence-model faces and 88 difference-model faces. There was no overlap in the face images between the two versions of the study.

The 175 chosen stimuli were presented in random order to each participant. For each stimulus, the question asked was "What is the gender of this person?" presented with two options: male or female. Left and right arrow keys were used to indicate one or the other gender, and the gender–key mapping was counterbalanced. Before the experiment began, each participant was told to rely on gut instinct, not to spend too much time on each face, and that there were no right or wrong answers. Participants were given unlimited time.

To assess intrarater reliability, we added 25 repeated trials randomly chosen from the first 175 trials in each study, bringing the total number of trials to 200. As we did in the previous experiments, we excluded the responses of participants with test-retest reliability less than or equal to 0: 1 participant. We recruited an additional participant so that we had 30 participants with test-retest reliability greater than 0.

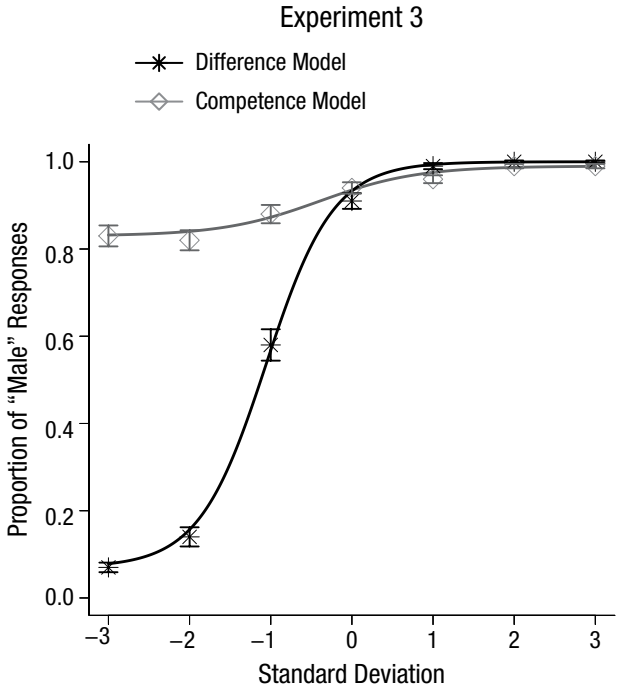

Fig. 4. Results of Experiment 3. The mean proportion of "male" responses is shown as a function of the level of the difference-model manipulation and the competence-model manipulation. Sigmoid functions were fitted for the response averaged across participants. Error bars denote standard errors.

Results

Overall, faces were more likely to be categorized as male than as female: The proportion of "male" responses to all faces (n = 350) averaged across participants was significantly higher than .5 (M = .79, SD = .31), $t(349)$ = 17.55, $p < .001$. This may be mainly attributed to the fact that the faces were bald, which creates a strong bias to perceive faces as male. Nevertheless, as shown in Figure 4, as the competence-manipulation level increased in both models, the categorization of faces as male increased, too.

To test whether the perceived gender of faces tracked the model manipulations, we conducted a 7 (manipulation level) × 2 (model type) repeated measures analysis of variance on the proportion of "male" responses for each face. This analysis found that perceived gender varied as a function of both manipulation level and the type of model, as well as their interaction. First, faces were more likely to be categorized as male when they were manipulated to look more competent, irrespective of the model type, as indicated by a main effect of impression-manipulation level, $F(6, 144) = 464.32$, $p < .001$, $\eta^2 = .88$. Second, faces were more likely to be categorized as male when they were manipulated by the competence model than when they were manipulated by the difference model, as indicated by a main effect of impression model, $F(1, 24) = 796.55$, $p < .001$, $\eta^2 = .69$. Third, the difference model led to a much larger difference in the proportion of "male" categorization responses as a function of the manipulation level than the competence model did, as indicated by the interaction effect between manipulation level and impression model, $F(6, 144) = 291.93$, $p < .001$, $\eta^2 = .78$.

This interaction effect reveals that the two models had differential effects on gender perception. When faces were varied by the standard competence model, most of the faces were likely to be categorized as male, despite the main effect of the manipulation level. This bias to perceive the faces as male could be attributed to the fact that they were all bald. However, once the attractiveness of the faces was subtracted from the faces manipulated by the competence model, as in the faces manipulated by the difference model, gender-categorization responses changed dramatically. Whereas faces manipulated to be perceived as competent (but not more attractive) were categorized as male, faces manipulated to be perceived as less competent (but not less attractive) were categorized as female (we obtained similar results using a competence model orthogonal to the attractiveness model; see Experiment S3 in the Supplemental Material). This effect shows that after the positive covariance between attractiveness and competence impressions is visually removed, the variance in the masculinity cues becomes much more prominent in the faces.

Experiment 4

The results of Experiments 2 and 3 suggest gender biases in competence impressions. However, because we used synthetic faces, which tend to be categorized as male because of lack of strong gender cues such as hair (Experiment 3), it is unclear whether masculinity cues influence impressions of male and female faces in the same way. Experiment 4 had two objectives: to extend our findings to real-life images of faces, which are unambiguously categorized as male or female, and to test for potentially differential effects of masculinity cues on competence impressions of male and female faces.

We manipulated photo-realistic male and female faces using the difference model. For male faces, we expected a monotonic increase in competence impressions when faces were manipulated to be more masculine. For female faces, the predictions were less clear, because both empirical findings and computational work show that female faces are evaluated more negatively if their appearance is counterstereotypical (Oh, Dotsch, Porter, & Todorov, 2018; Sutherland, Young, Mootz, & Oldmeadow, 2015). Given the strong gender stereotypes, it is possible that masculinity cues would have no effect or even a negative effect on competence impressions of female faces. Alternatively, whereas weak masculinity cues may increase these impressions, strong masculinity cues may decrease them.

Method

Participants. Two hundred sixty-eight MTurk workers (153 men, 115 women; age: M = 38.24 years, range = 21–71) were recruited and participated for payment. A power analysis using G*Power 3.1.9 indicated that a sample size of at least 122 participants per condition would afford 90% power to detect a small effect (R^2 = .08) of impression-manipulation levels in participant-level regressions, which was observed in Experiment 1b for the competence rating.

Materials. To apply a computational model to real-life face images, we transformed real face images using PsychoMorph (Tiddeman, Burt, & Perrett, 2001). First, we selected 10 male and 10 female real-life face images of self-identified Caucasian individuals (see Fig. S8 in the Supplemental Material) from a standardized face image set (DeBruine & Jones, 2017). Then, we created synthetic faces that represented extreme faces (−3 and 3 *SD*, shown in the bottom row of Fig. 1) generated by the difference model. Next, we manipulated the 20 real-life face images along the continuum of the difference between the two extreme face images in shape, color, and texture. The transformation procedure allowed us to generate photo-realistic images that varied in facial information captured by the difference model (see Fig. 5). On the most extreme ends, each face image was transformed 25% away from the original face. The manipulation magnitude was identical for the intervals between the four facial variations: The final face images were transformed −25.00%, −8.33%, 8.33%, and 25.00% away from the initial images. The complete set of stimuli consisted of 80 face images (2 face genders × 10 identities × 4 manipulation levels).

Procedure. Each participant was randomly assigned to make judgments of competence of either male faces (male-rating condition) or female faces (female-rating condition). The stimuli were presented in random order to each participant with the constraint that face images generated from the same original face identity were never consecutively shown. For each stimulus, the question asked was "How competent is this person?" which participants responded to using a 9-point scale ranging from 1 (*not at all competent*) to 9 (*extremely competent*). Before the experiment began, each participant was told to rely on gut instinct, not to spend too much time on each face, and that there were no right or wrong answers. Participants were given unlimited time to respond.

To assess intrarater reliability, we added 10 repeated trials randomly chosen from the first 40 trials in each study, bringing the total number of trials to 50. Using the ratings from the 10 repeated trial pairs, we calculated a correlational coefficient as a measure of test-retest reliability of each participant. We excluded from further analyses the responses of any participants with test-retest reliability less than or equal to 0: 13 participants in the male-rating condition and 5 participants in the female-rating condition. We recruited additional participants so that we had 125 participants with test-retest reliability greater than 0 per face gender. The interrater reliabilities of the impression ratings were high (male faces: α = .95, female faces: α = .97).

Results

Linear and quadratic regression models were fitted for the competence ratings to test whether competence impressions tracked the difference-model manipulation. For the regressions, the ratings were averaged across face identities (participant-level analysis, n = 125 per gender; see Fig. 6). The linear model explained a significant amount of variance in the ratings of male faces but not in the ratings of female faces—male faces: R^2 = .02, $F(1, 498) = 11.94$, $p < .001$; female faces: $R^2 < .01$, $F(1, 498) = 2.41$, $p = .121$. The quadratic model explained a significant amount of variance in the ratings of faces of both genders—male faces: $R^2 = .02$, $F(2, 497) = 6.04$,

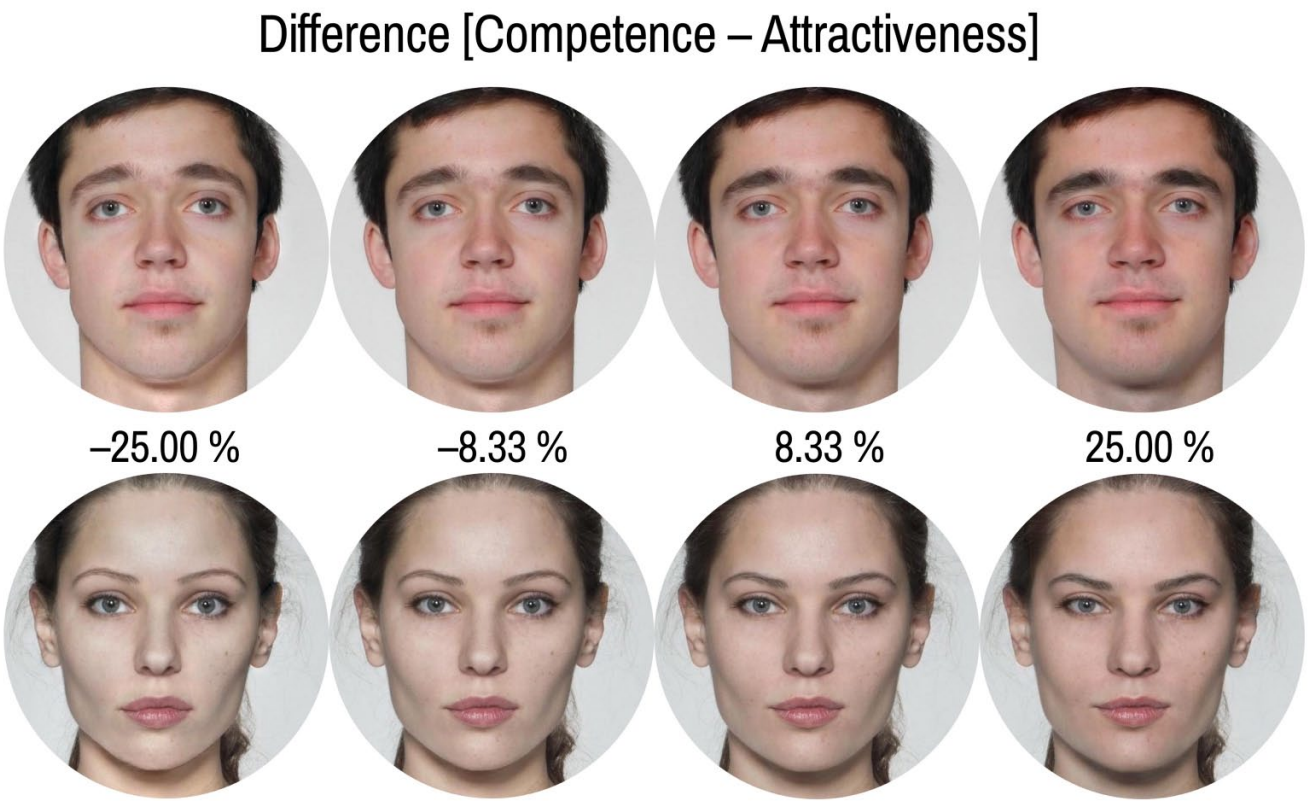

Fig. 5. Example real-life face images manipulated by the difference model for Experiment 4. As the manipulation increases, the faces are perceived as more competent. The unit of manipulation represents the extent to which the shape and reflectance of each original face image were transformed toward or away from the extreme ends of the model (i.e., –3 and 3 *SD* faces in Fig. 1, bottom).

p = .003; female faces: R^2 = .02, $F(2, 497) = 4.29$, p = .014. The quadratic fit was better than the linear fit for female faces, $F(1, 497) = 6.15$, p = .014, but not for male faces, $F(1, 497) = 0.16$, p = .690.

Although the observed effect sizes of the model manipulation were significant, they were relatively small. This finding may be due to several factors: The difference model did not directly manipulate facial-attractiveness cues, which play a big role in competence impressions; the difference model was derived from the ratings of synthetic, not real-life, face images and therefore might have had a smaller effect when applied to real-life faces; real-life face images are more distinctive than synthetic face images across face identities, adding to unexplained variance; and participants may have regarded face images originating from the same identity as identical and therefore provided the same rating across these images before closely examining the facial cues.

Nevertheless, the results show that when facial cues of competence impressions were enhanced using the difference model, competence impressions increased, although the effects were different for male and female faces. Whereas male faces became more competent looking through the increasing manipulation levels (see Fig. 6), female faces became more competent looking only up to a point, after which they became less competent looking (as shown in the significant quadratic fit of the ratings; see Fig. 6). These results show that (a) perceived competence of real-life male and female faces can be meaningfully altered, controlling for the halo effect of attractiveness; (b) male faces receive incremental benefit as masculinity and confidence cues increase; and (c) female faces receive benefit to a certain extent as masculinity and confidence cues increase and begin to appear as not more competent when these cues become too strong (we obtained similar results using a competence model orthogonal to the attractiveness model; see Experiment S4 in the Supplemental Material). The latter finding is consistent with prior research showing that counterstereotypical (e.g., dominant) female faces are evaluated negatively (e.g., Oh et al., 2018; Sutherland et al., 2015) and theoretical frameworks positing that women are evaluated positively only when they fit narrowly defined female gender norms (e.g., Glick & Fiske, 1996).

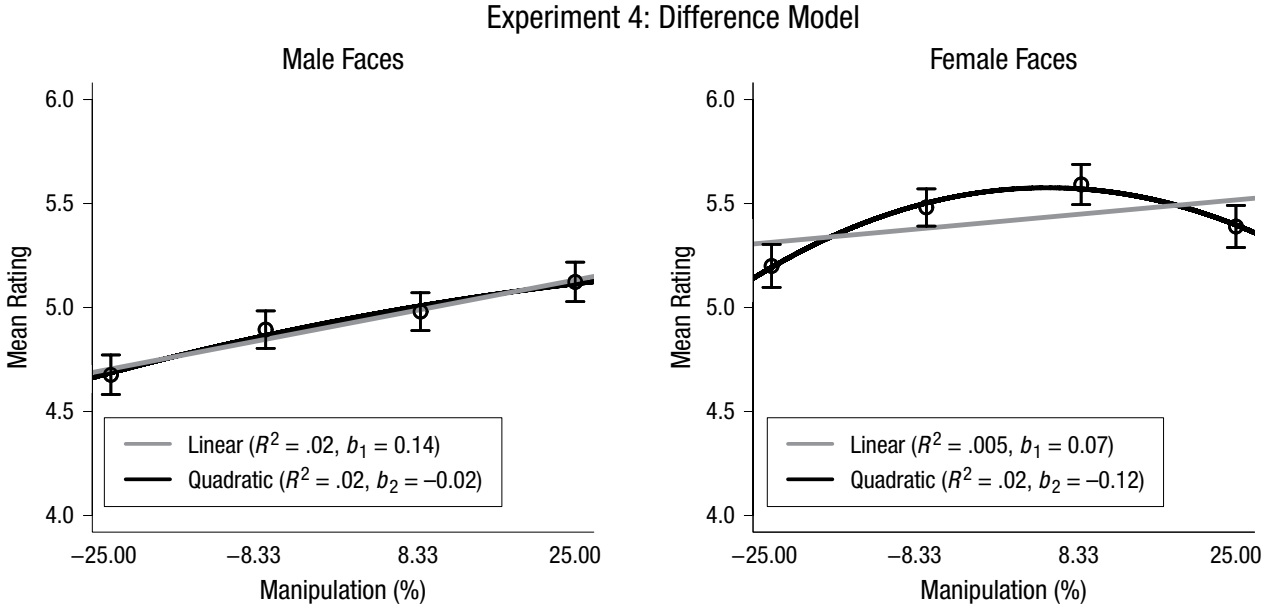

Fig. 6. Results of Experiment 4. The mean competence ratings of real-life male faces (left) and real-life female faces (right) as a function of the level of the difference-model manipulation. For each impression, fits of the linear and quadratic models are shown for the mean rating, averaged across faces. Error bars denote standard errors.

Female faces were, on average, rated as more competent (M = 5.42, SD = 1.09) than male faces (M = 4.92, SD = 1.04), $t(995.62) = 7.35$, $p < .001$ (see Fig. 6). This may be due to several factors: Female faces are, on average, more attractive than male faces because of their feminine facial features (e.g., Said & Todorov, 2011), and attractiveness is an ingredient of competence impressions (as shown in Experiments 1–3, without a proper control for attractiveness, it is not readily apparent that masculine facial features contribute to competence impressions); each participant was exposed to either only male faces or only female faces, and the two groups of participants might have adopted different psychological scales for the rating task.

General Discussion

We set out to identify the visual components of competence impressions from faces using data-driven computational models of impressions. To begin with, we showed that competence impressions are naturally correlated with facial attractiveness (Experiment 1a). When the variance in competence impressions could not be attributed to facial attractiveness (Experiment 1b), more competent-looking faces were perceived as more confident and masculine (Experiment 2) and were more likely to be categorized as male (Experiment 3). In addition, the confidence and masculinity cues related to competence impressions increased perceived competence in real-life images of male faces, but only to a limited extent (and with a backfiring effect) in female faces (Experiment 4).

In sum, we found that cues for attractiveness, confidence, and masculinity are at the core of face-based competence impressions, consistent with the idea that competence impressions have multiple components. Crucially, the findings suggest that there are strong gender biases that shape competence impressions. Specifically, men are more likely to be perceived as competent, an effect that was not readily apparent from a model of competence impressions that did not control for facial attractiveness. When we increased the masculinity cues in real-life face images, these cues benefited men's impressions of competence but benefited women's impressions only up to a point (Experiment 4). When the masculinity cues were too strong, women were perceived as less competent, an effect that is largely attributed to strong biases to negatively evaluate women with counterstereotypical looks (e.g., Oh et al., 2018; Sutherland et al., 2015). These findings have significant implications for social-perception research and social fairness.

First, the results illustrate a useful way to overcome natural confounds in visual social perception (Todorov et al., 2013). Prior research (e.g., Dion et al., 1972; Eagly et al., 1991; Olivola & Todorov, 2010; Todorov et al., 2013) and Experiment 1a showed a strong relationship between judgments of attractiveness and competence. Generally, attractiveness correlates with many other social judgments (Dion et al., 1972; Graham et al., 2017;

Todorov et al., 2013; Webster & Driskell, 2015), making it difficult to test whether effects of various social impressions on behaviors are due to specific impressions, such as trustworthiness and competence, or the halo effect of attractiveness. However, within the present computational framework, it is straightforward to control for facial attractiveness: We illustrated this method by simply subtracting the model of attractiveness from the competence model. This is the strongest manipulation of controlling for attractiveness, because it forces the two judgments, as shown in Experiment 1b, to be negatively or not correlated. It is also possible to orthogonalize two different models (see Fig. S9 in the Supplemental Material), but the orthogonality between the models in the face-model space does not guarantee that judgments of faces generated by the two models would be orthogonal. In fact, when empirically tested, the orthogonal model could not eliminate the attractiveness confound from competence impressions (see the supplemental experiments in the Supplemental Material). In addition to controlling for undesirable confounds, these methods have the potential to reveal ingredients of impressions that are not readily discernible from the models of these impressions. As shown in Experiments 2 through 4, after we controlled for attractiveness, it was readily apparent and validated that competence impressions rely on masculinity cues.

Second, the current findings are consistent with a body of literature in psychology and gender studies that shows a strong link between competence impressions and maleness. People report that both other people ("society") and they themselves think that men are more competent than women (Bem, 1974; Broverman et al., 1972; Spence et al., 1975). This gender bias in perceived competence emerges early in development. Girls as young as 6 years old, unlike boys of the same age, think that their gender, on average, lacks a high level of intellectual competence (Bian, Leslie, & Cimpian, 2017). These biased perceptions most likely exacerbate gender inequality in various settings: Women are more likely to be discriminated against in a professional environment (Hagen & Kahn, 1975; Rudman & Phelan, 2008) and are more likely to face obstacles when entering or staying in a field that requires intellectual competence (Leslie, Cimpian, Meyer, & Freeland, 2015; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012).

It is noteworthy that certain feminine facial features, especially feminine face shape, make both male and female faces more attractive (Perrett et al., 1998; Rhodes et al., 2000; Said & Todorov, 2011; but see Rhodes, 2006). Thus, one may think that women, especially those with attractive or feminine facial looks, may benefit from the halo effect of attractiveness when their competence is judged by a stranger. However, the current findings suggest that masculinity cues play a crucial role in competence impressions, hurting women to the extent that they possess feminine (i.e., less masculine) looks. On the other hand, when women possess strong masculine looks, they may be discriminated against because of the counterstereotypical nature of these looks (Oh et al., 2018; Sutherland et al., 2015).

The gender biases in competence impressions are masked because the attractiveness caused by feminine facial features contributes to competence impressions. By effectively removing the influence of attractiveness on competence impressions, the current research uncovered strong gender biases in these impressions. These biases are particularly alarming because intuitive competence judgments have powerful effects on leadership selection (Antonakis & Eubanks, 2017). Ideally, facial components related to gender should not have any effect on how we form first impressions of competence. Our findings, on the contrary, suggest that gender-related facial cues play a key role in competence impressions, potentially creating hostile environments for women.

Action Editor

Alice J. O'Toole served as action editor for this article.

Author Contributions

A. Todorov developed the study concept. All the authors contributed to the study design. D. Oh and E. A. Buck collected and analyzed the data. All the authors wrote the manuscript and approved the final manuscript for submission.

ORCID iD

DongWon Oh https://orcid.org/0000-0002-2105-3756

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618813092

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at osf.io/ygzx3 and osf.io/86kfq, respectively. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618813092. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

References

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

Antonakis, J., & Eubanks, D. L. (2017). Looking leadership in the face. *Current Directions in Psychological Science*, *26*, 270–275. doi:10.1177/0963721417705888

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

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42*, 155–162.

Bian, L., Leslie, S.-J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*, 389–391. doi:10.1126/science.aah6524

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*(2), 59–78. doi:10.1111/j.1540-4560.1972.tb00018.x

DeBruine, L., & Jones, B. (2017). *Face Research Lab London Set*. Retrieved from Figshare: https://figshare.com/articles/Face_Research_Lab_London_Set/5047666

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*, 285–290.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but . . . : A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*, 109–128.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi:10.3758/BF03193146

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. doi:10.1167/14.1.28

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*, 491–512.

Graham, J. R., Harvey, C. R., & Puri, M. (2017). A corporate beauty contest. *Management Science*, *63*, 3044–3056. doi:10.1287/mnsc.2016.2484

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

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

Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology*, *29*, 299–304. doi:10.1037/h0036018

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. doi:10.1111/j.1540-5907.2011.00511.x

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*, 262–265. doi:10.1126/science.1261375

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, USA*, *109*, 16474–16479. doi:10.1073/pnas.1211286109

Oh, D., Dotsch, R., Porter, J. M., & Todorov, A. T. (2018). *Gender biases in impressions from faces: Empirical studies and computational models*. Retrieved from PsyArXiv: https://psyarxiv.com/fxvcu/

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

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

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

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, *57*, 199–226. doi: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. doi:10.1348/000712600161718

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

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

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.

Stoker, J. I., Garretsen, H., & Spreeuwers, L. J. (2016). The facial appearance of CEOs: Faces signal selection but not performance. *PLOS ONE*, *11*(7), Article e0159950. doi:10.1371/journal.pone.0159950

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. doi:10.1111/bjop.12085

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*, 25–29.

Tiddeman, B., Burt, M., & Perrett, D. I. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications*, *21*(4), 42–50. doi:10.1109/38.946630

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

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

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

Todorov, A. T., 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.

Todorov, A. T., & Oosterhof, N. N. (2011). Modeling social perception of faces. *IEEE Signal Processing Magazine*, *28*(2), 117–122. doi:10.1109/MSP.2010.940006

Webster, J., Jr., & Driskell, J., Jr. (2015). Beauty as status. *American Journal of Sociology*, *89*, 140–165. doi:10.1086/227836

Williams, J. E., & Best, D. L. (1990). *Measuring sex stereotypes: A multination study*. Newbury Park, CA: SAGE.

Wyatt, M., & Silvester, J. (2018). Do voters get it right? A test of the ascription-actuality trait theory of leadership with political elites. *The Leadership Quarterly*, *29*, 609–621. doi:10.1016/j.leaqua.2018.02.001

Zellner, A. (1963). Estimators for seemingly unrelated regression equations: Some exact finite sample results. *Journal of the American Statistical Association*, *58*, 977–992. doi:10.1080/01621459.1963.10480681