Abstract

Many studies have investigated whether facial averageness, masculinity, and similarity are associated with facial attractiveness. These studies have relied on ratings of images of real or digitally morphed faces. It is important to establish whether past findings translate to real-life, face-to-face evaluations of potential partners; lack of effects in this context would cast doubt on the evolutionary relevance of previous findings. Further, previous studies have not considered that, by definition, faces that are more similar to the average face (i.e. higher in averageness) tend to be more similar to raters' faces. Therefore, these image-rating studies, which separately found that averageness and (in some cases) similarity are attractive, are confounded. To address these issues, we conducted a laboratory-based speed-dating study of 682 participants with objectively measured facial traits, where opposite-sex pairs rated each other on facial attractiveness and prosociality. We found that facial attractiveness was predicted separately by averageness and by similarity (to the rater), but with both variables in the same model, neither uniquely predicted attractiveness. Similarity, but not averageness, predicted prosociality ratings. Facial masculinity was positively and negatively associated with facial attractiveness ratings of men and women, respectively. These results confirm, in real-life interactions, some key findings from image-rating studies but raise questions about others, notably the attractiveness of facial averageness.

Keywords: averageness, similarity, masculinity, morphometrics, kinship, mating, relationship formation, partner selection

Funding**:** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability: The data is not publicly available and will be provided upon request.

1. Introduction

Faces are not only essential for identifying individuals, but they are also an important contributor to evaluations of physical attractiveness (Sidari et al., 2020). Given that there is a cross-cultural consensus on what kind of faces are considered attractive (Cunningham et al., 1995; Perrett et al., 1994; G. Rhodes et al., 2001) and that preferences for these attractive faces emerge early in development (Geldart et al., 1999; Langlois et al., 1991; Langlois et al., 1987; Rubenstein et al., 1999), there is likely a genetic basis for what kind of faces we find attractive. Therefore, preferences for certain facial features potentially evolved due to fitness benefits signalled by those features.

Studies that have sought to determine what facial features we may have evolved to find attractive have yielded conflicting evidence (e.g. Lee, Mitchem, et al., 2014; Penton-Voak & Chen, 2004; Rhodes, 2006). Despite humans having relied (until most recently) on face-to-face interactions to interact and form judgments about another person, facial attractiveness studies thus far have relied exclusively on participants rating the attractiveness of images presented on screens or on paper. Two-dimensional images do not capture the range of expressions and angles that are experienced when viewing or interacting with another person. There have been few studies that use objective facial measures to predict in-person evaluations (e.g. Hill et al. (2013); Valentine et al. (2014)) and none of these have evaluated attractiveness, so it is unclear whether past attractiveness research based on facial images generalises to real-time, real-life interactions with a person who in addition to a face, has a body and a personality. Indeed, it could be that the correlates of attractiveness identified using isolated facial images wash out when mixed with all the other variables involved in real life interactions.

In this study we assess the impact of facial masculinity/femininity, averageness, and similarity on attractiveness as perceived in real-life interactions; first though, we give a brief overview of existing facial attractiveness studies concerning these characteristics.

1.1. Facial Masculinity/Femininity

For women, studies have consistently shown a positive association between facial femininity and rated attractiveness across both manipulated and natural photos of subjects (Alharbi et al., 2020; Foo et al., 2017; Lee, Dubbs, et al., 2014; Rhodes, 2006). For men, findings vary from negative to positive associations between facial masculinity and rated attractiveness depending on the nature of the stimuli used (Foo et al., 2017; Lee, Mitchem, et al., 2014; Penton-Voak & Chen, 2004; Rhodes, 2006).

Facial masculinity/femininity has typically been assessed using raters' subjective impressions of how feminine or masculine the subject of the photo is according to their personal interpretation of the terms (e.g. Foo et al., 2017; Scott et al., 2010). In this approach, a rater's definitions of masculinity and femininity may be influenced by social expectations, contaminating analyses aimed at analysing the biological or structural aspects of faces. Additionally, in many cultures, conforming to one's social role (i.e. being a masculine man or feminine women) is considered a positive trait. Therefore, subjective ratings of facial masculinity/femininity may be subject to a halo effect, whereby raters may assign spuriously higher masculinity scores to attractive male faces, for example. Improvements in technology, though, have allowed for the use of objective measures to quantify the degree of masculinity/femininity of natural images of individuals' faces. These measures use data-driven statistical techniques, primarily linear discriminant analysis. This method linearly combines information from facial landmark coordinates in a way that best categorises faces on sex, assigning a masculinity/femininity score for each face. This objective approach addresses the limitations of subjective measures by removing the possibility of rater bias.

1.2. Facial Averageness

Facial averageness is defined by the extent that an individual resembles the average of faces from a given sample, usually of their own sex. Conversely, facial distinctiveness is the extent that an individual's face is different from the average. From an evolutionary perspective, facial averageness is thought to be an indicator of underlying genetic quality, as evolutionary pressures may act against extremes of a trait in favour of the most common or average features (Dobzhansky, 1982); deviations from this average may be caused by genetic mutations or developmental perturbations (e.g. infectious disease). Indeed, there has been a general consensus across studies, irrespective of stimuli type and the measure of averageness, that more average faces are rated as more attractive than more distinctive faces (Cai et al., 2019; Foo et al., 2017; Lee et al., 2016; Light et al., 2016; Rhodes, 2006).

Even though averageness is a mathematical quantity, past studies have primarily used non-mathematical techniques to quantify or manipulate levels of facial averageness. Some studies manipulate the facial averageness of stimuli by increasing the number of times an original image is morphed with other individuals; the higher the number of image-morphs, the higher the facial averageness of the final composite. However, realistic facial features such as skin texture and facial hair can be lost in this process. Other studies quantify facial averageness with subjective ratings, whereby raters are asked how distinctive a given image of a face is, and the resulting rating is reverse-scored to obtain an averageness rating (Foo et al., 2017; Light et al., 2016; Gillian Rhodes et al., 2001). A limitation of this measure is that interpretations of distinctiveness can be conflated with what is unfamiliar, resulting in facial averageness being confounded with familiarity (Light et al., 2016). With methodological advancements, we are able to address the shortcomings of prior methods by using objective measures of facial averageness. We can obtain an averageness score by calculating the extent that the landmarks of an individual face are similar to the landmarks from the average of a sample of faces (e.g. Lee et al., 2016).

1.3. Facial Similarity and Attractiveness

Humans mate assortatively across a variety of different traits; that is, they partner up with people who are similar to themselves (Zietsch et al., 2011). There has been mixed evidence that couples are similar with respect to faces (Abel & Kruger, 2011; Alvarez & Jaffe, 2004; DeBruine, 2004; Griffiths & Kunz, 1973; Hinsz, 2016). If we assume that couples do have similar faces, it is still unclear why this would occur. Some have proposed that long-term shared lifestyles result in a convergence of physical likeness, but there is mixed evidence regarding this possibility (Tea-Makorn & Kosinski, 2020; Zajonc et al., 1987). Alternatively, facial similarity in couples could be the result of a preference for facial similarity in a romantic partner.

However, there is also mixed evidence for an association between facial similarity and ratings of attractiveness in opposite-sex faces. Such studies have used stimuli consisting of manipulated images of opposite-sex faces that are morphed to resemble the participant. Using this method, Penton-Voak and Chen (2004) found that females rated self-similar male morphs more attractive than males morphed with other faces. On the other hand, both DeBruine (2004) and Lindova et al. (2016) found no such association using the same method with both male and female participants.

1.4. Facial Similarity and Prosociality

An alternative explanation for the preference for similar faces is that this preference is not driven by the attractiveness of similar faces per se, but that similar faces may signal kinship (DeBruine, 2002; Hansen et al., 2020). While cues of kinship can decrease sexual desirability due to the costs of inbreeding (Bittles & Neel, 1994), the same cues can increase prosocial behaviour by promoting inclusive fitness benefits (Hamilton, 1964). There is also evidence that positive assortative mating can result in net positive reproductive benefits, as a small degree of relatedness is associated with higher fertility (Helgason et al., 2008; Weller & Santos, 2013).

Kinship cues not only have relevance in potential partner choice, but also in social situations. Kinship can signal in-group status and therefore elicit prosocial behaviours and attitudes. DeBruine (2002) found that participants who faced computer game opponents who facially resembled themselves were more willing to trust the opponent. Indeed, DeBruine (2005) found that while individuals evaluated similar facial morphs as more trustworthy than dissimilar facial morphs, similar faces were judged as no more attractive than dissimilar faces. On the other hand, Lindova et al. (2016) found no effect of opposite-sex facial similarity on perceptions of trustworthiness. Although the evidence for in-group biases is strong (e.g. Everett et al., 2015; Tajfel, 1970), it is unclear whether the purported effects of facial similarity on attractiveness and prosociality ratings would prevail in a real-life context.

1.5. General Limitations of Past Studies

In addition to mixed findings with regard to the effects of facial similarity, studies in this area have not accounted for the possibility that perceptions or calculations of facial similarity are confounded by the opposite-sex nature of the stimuli. For example, the similarity of a female participant to an image of a male will tend to depend on how similar the participant’s face is to male faces in general (i.e. how masculine her face is) as well as how similar it is to that particular male. On average, a female participant’s face would be more similar to a feminine male face than a typical or less feminine male face, and this could influence the attractiveness of facial similarity. In our study, we account for this effect statistically with a new variable computed by subtracting the average male and female faces from male and female participants, respectively, before calculating dyad similarity.

Additionally, no study has controlled for facial averageness when evaluating the relationship between facial similarity and attractiveness. Penton-Voak et al. (1999) proposed that a target who is more facially average will tend to be both more attractive (because of the attractiveness of facial averageness) and more similar to the participant (because average faces are by definition more similar to other faces), but this idea has not been explicitly investigated. This confound could generate a spurious association between similarity and attractiveness if averageness were not controlled for.

1.6. Present Study

Here, we use a speed-dating paradigm to gather evaluations of facial attractiveness (and trustworthiness) in live interactions, and test their association with objective measures of facial masculinity/femininity, averageness, and similarity. We use a large sample (N = 682) to enable precise estimation of subtle effects, and we control for potential confounds that have not previously been accounted for. Thus, this study aims to provide a high-powered, robust, and wide-ranging examination of correlates of facial attractiveness in a live-interaction context.

2. Method

2. 1. Participants

Participants were 682 (344 female) first-year psychology students at the [redacted](female ages: M =19.17, SD = 2.83; male ages: M =19.94, SD = 3.05). There were 437 Caucasian, 220 Asian, 9 Middle Eastern, 7 African, 7 Latino, and 2 Pacific Islander participants (note: nine participants did not specify any ethnicity, so they were visually classified by three of the authors). There were 2285 speed-dating interactions in total; 1188 consisted of interactions that were ethnically concordant (i.e. participants interacted with a partner of the same ethnicity), and 1097 consisted of interactions that were ethnically discordant (i.e. participants interacted with a partner of a different ethnicity).

Participants were recruited in exchange for one credit towards a research participation course, or through word of mouth. Participants were eligible if English was their first language, they were single, heterosexual, and if they were open to answering sensitive questions about topics such as their sexual history. Participation was voluntary and participants were repeatedly notified that they could withdraw at any time without consequence. This speed-dating study was a part of an ongoing study running from 2010 to 2019. However, data from the 2014, 2015, 2018, and 2019 iterations were eligible for use in this study as photographs of participants were only available in these years.

2.2. Materials

Participants responded to various questionnaires that were a part of a larger study investigating attraction. Measures relevant to the current study are described below.

2. 2. 1. Demographics

Measures of participant demographics included age, sex, ethnicity (Caucasian, Asian, African, or Other), and English proficiency.

2. 2. 2. Speed-dating ratings

Participants completed a questionnaire about each partner with whom they had a speed-dating interaction. Participants were asked ‘Please rate this partner on the following statements below’. Facial attractiveness was measured using responses to the statement ‘They are facially attractive’. Prosociality was measured using the statement ‘They are kind and understanding’; we note that we were not able to measure prosociality more explicitly as our measures were designed for a long-running speed-dating study without regard to the question about prosociality considered here. All items were rated on a 7-point scale (1 = *Well Below Average* to 7 = *Well Above Average*).

2. 3. Procedure

2. 3. 1. Pre-date

The study was conducted in a laboratory with two large rooms, each containing two speed-dating stations. Each station contained a table and two chairs so that participants could sit opposite each other; these stations were partitioned with room dividers to ensure privacy during each interaction. Males and females were seated in separate rooms upon arrival to ensure minimal communication between members of each sex before the study. There was a minimum of two and a maximum of five participants of each sex per speed-dating session (depending on sign-up and attendance rates). All participants were given a photo consent form and an information sheet about the study. Participants had the option to withdraw consent while still partaking in the speed-dating study. Once participants had signed the consent form and read the information sheet, they completed the questionnaire on their iPad which included demographic questions.

2. 3. 2. Speed-date

Participants were verbally informed about the activities involved in the study; they had three minutes to get to know a person of the opposite sex, and they would answer questions about their interaction on their iPads. We then rearranged participants so one of each sex was at each station, and participants without a partner sat by themselves during the current interaction.

A bell was rung to commence and conclude the interaction. We then advised participants to stop talking and to fill out the questionnaire regarding their interaction. If there was a participant with no partner in that interaction round, they skipped the partner ratings questionnaire. This process was repeated until all possible opposite-sex dyads interacted.

2. 3. 3. Post-date

Participants completed the rest of the questionnaire. We then verbally debriefed participants on the purpose of the experiment and handed them a debriefing sheet.

2. 3. 4. Facial Photographs

Participants were asked to stand in front of a blank wall or door and to look straight ahead with a neutral expression before their facial photos were taken using a Canon IXUS 700. To keep images consistent, we used the flash and no zoom along with consistent indoor lighting.

2. 4. Measures

2. 4. 1. Landmarking and Photograph Processing

A total of 688 photos were obtained, though six of these participants did not participate in speed dating due to insufficient attendance at their sessions. All photos were landmarked using Webmorph (DeBruine, 2018) according to a 28-landmark template, where each landmark is defined by a specific facial location (e.g. left pupil) that is represented by an x and y coordinate. The location of certain facial points (such as the hairline) may be obscured by hair or head orientation, affecting the accuracy of landmark placement. To address this, we used sliding semilandmarks to account for homologous curves and outlines in our Procrustes analysis. An example of both landmarks and semilandmarks are depicted in Figure 1.

Figure 1.

Example of 28 landmarks (green crosses) and semilandmarks (blue lines) placed on a sample composite photo from (DeBruine, 2016).

The landmarking of photographs from 2014 and 2015 was performed by a research student, and the same research student was assisted by a co-author for the landmarking of the 2018 photographs. The two sets of landmarks from the 2018 photographs were averaged to give one set of landmarks. The 2019 photographs were landmarked by two other research students. The first author checked the statistical and visual agreement of the 2019 landmarks from the two research students, and while statistically, the agreement was very high, there were occasional but clear errors in one of the coder's landmarks, so we used the other coder's landmarks for analysis.

We used geometric morphometrics to analyse the shape information via the facial landmarks and semilandmarks using the "geomorph" package in R (Adams et al., 2022; Baken et al., 2021). A generalised Procrustes analysis was used to extract shape information from the landmark coordinates. This procedure standardises all faces by removing the effects of translation, rotation, and scaling, with only shape information remaining in the form of Procrustes coordinates. The semilandmarks allowed the Procrustes analysis to align landmarks while accommodating more accurately for possible inaccuracies in landmark placement along these semilandmark curves (Adams et al., 2004; Bardua et al., 2019). These aligned coordinates were then transformed into shape variables via principal components analysis. In this step, we obtained 54 principal components, accounting for 100% of total variance. For full details of generalised Procrustes analysis and shape analysis, see Zelditch et al. (2004). We used principal components for the facial masculinity calculations, but we used the aligned landmarks for the facial averageness and similarity calculations using the package "facefuns" (Holzleitner & DeBruine, 2020).

2. 4. 2. Facial Averageness Scores

Facial averageness measures how close a person's face is to the geometric average face of their sex. Facial distinctiveness (the opposite of averageness) was given by the Euclidean distance (the square root of the sum of squared distances across each landmark) of each participant's face from the average face of their sex. The higher the Euclidean distance, the higher the extent that one's features differ from the average, and therefore the more distinctive the face is. For ease of interpretation, facial averageness scores were calculated by multiplying distinctiveness scores by negative 1.

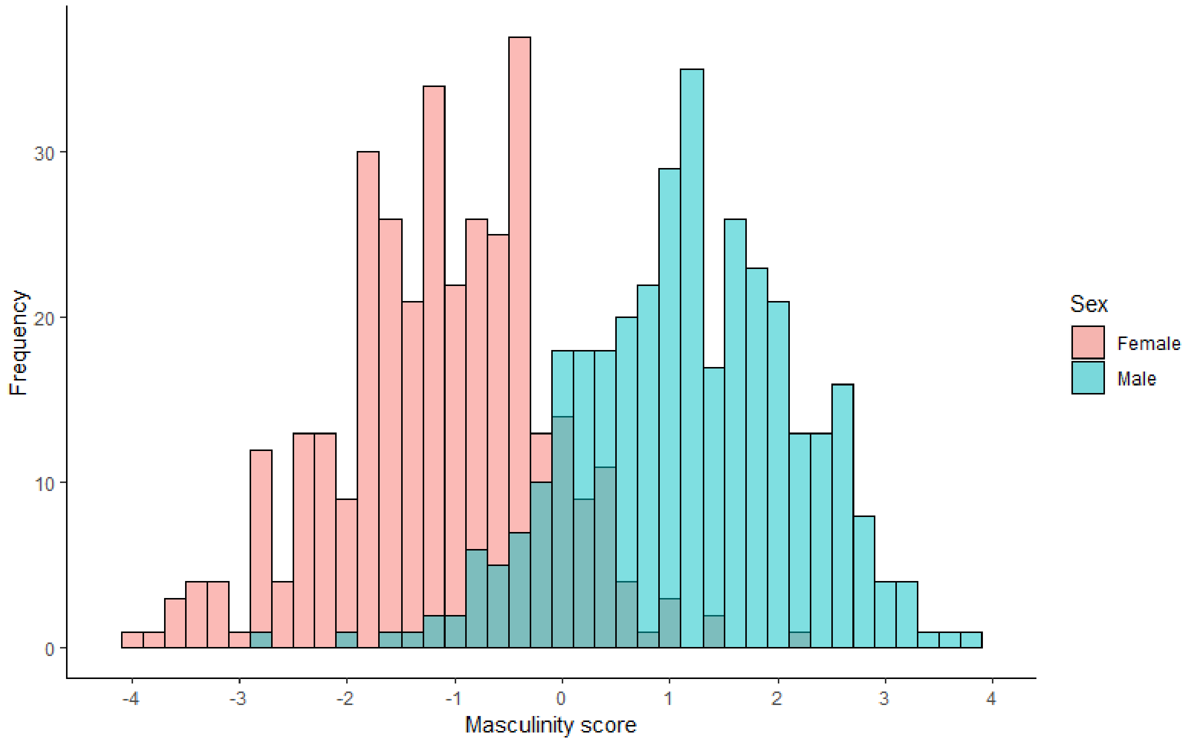
2. 4. 3. Masculinity Calculation

Linear discriminant analysis was used to calculate masculinity scores using the package "MASS" (Venables & Ripley, 2002) in 'R' (R Development Core Team, 2021). Linear discriminant analysis uses shape variables to fit a model that best discriminates males from females in the training sample (e.g. Lee, Mitchem, et al., 2014). The resulting linear discriminant function represents the female-male continuum (Figure 2).

We used all 688 landmarked images. Ten-fold cross validation was used to maximise the dataset for training and testing. Validation was used to minimise prediction bias (Hastie, 2009). The 10-fold cross validation was repeated 10 times, giving an average cross-validation accuracy of 84.3%. Cohen's κ accounts for the model correctly classifying the sex of the participant by chance. We obtained a κ value of 68.6%, – which demonstrates "substantial" reliability of the model (Landis & Koch, 1977). When the model was applied to the overall dataset, we found minimal overfitting as the average total accuracy was 89.2%, which was similar to the cross-validation accuracy (84.3%).

Figure 2.

Frequency distribution of objective masculinity scores from the linear discriminant analysis for males (M = 1.15, SD = 1.03) and females (M = -1.15, SD = 0.98).



2. 4. 4. Raw Facial Similarity Scores

Facial similarity scores were calculated for each speed-dating pair using the Euclidean distance between respective landmarks across the participant and partner’s faces. The Euclidean distance scores were multiplied by -1 so that a high score would indicate high facial similarity.

2. 4. 5. Sex-controlled Facial Similarity Scores

We created a sex-controlled facial similarity score to account for the different average faces of the two sexes. Instead of calculating the Euclidean distance between participant and partner faces, for the male and female participants, we subtracted the average male and female face, respectively, from the participant's own face, giving the participant's deviation from the average of their respective sex. We then calculated the Euclidean distance between the deviations of each member within a couple. The rationale for this procedure is that the raw facial similarity scores cause facial similarity to depend heavily on facial masculinity/femininity. For example, a woman whose face is maximally similar to a typical male partner will be exceptionally masculine (relative to other women), because her face will be the same shape as a typical male (i.e. her partner). In contrast, the sex-controlled facial similarity scores are not dependent on facial masculinity/femininity in this way – an opposite-sex couple will be maximally similar on this score when the male differs from the average male in the same ways as the female differs from the average female.

2. 5. Data Analyses

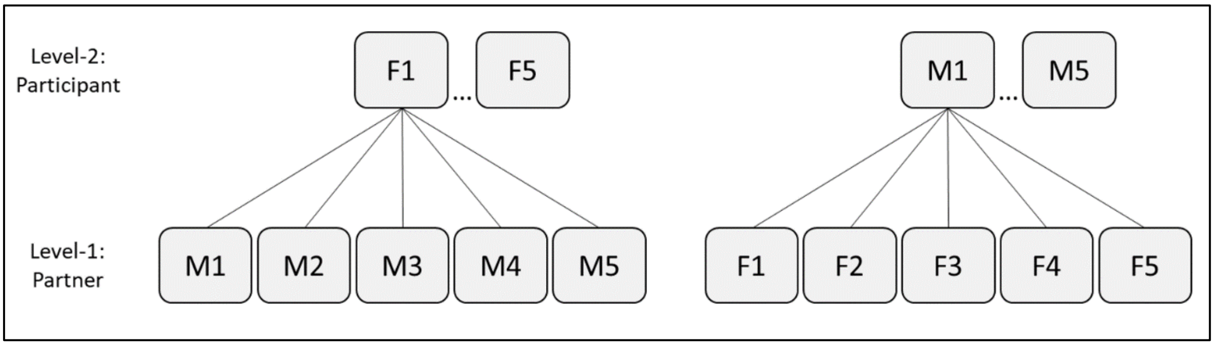
2. 5. 1. Hierarchical data

The design of the study is hierarchical in nature where the ratings of the partner are nested within the participant (

Figure 3). Here, it is necessary to use multilevel modelling as each partner's received ratings for *facial attractiveness*, and *kindness and understanding* are nested within the participant who gave the rating. Random intercepts for the year of the speed-dating session were included to account for variance that may be attributed to the annual inclusion and exclusion of questionnaires that were not relevant to the current study as well as possible cohort effects. Participant and partner random intercepts were included due to the dyadic nature of the data – these account for individual and partner differences. Data analysis was performed using the 'lme4' and 'lmerTest' packages (Bates et al., 2011; Kuznetsova et al., 2020) in 'R' (R Development Core Team, 2021). All quantitative variables were scaled such that the means and standard deviations of these variables were 0 and 1, respectively. We use γ to denote the standardised slope coefficients from our multilevel models.

Figure 3.

This model shows that within a speed-dating session, there are two levels where male partner ratings (M1 to M5) are nested within female participant F1 and that female partner ratings (F1 to F5) are nested within male participant M1.



2. 6. Ethics approval

This study was approved by the Low and Negligible Risk ethics committee at [redacted].

2. 7. Data

The data is not publicly available and will be provided upon request.

3. Results

We investigated the association between 1) facial averageness and rated facial attractiveness, 2) facial masculinity and rated facial attractiveness, and 3) facial similarity and rated kindness and understanding. We also conducted analyses using sex-controlled facial similarity in lieu of raw facial similarity where relevant. Participant sex was included as a covariate in all analyses.

3. 1. Preliminary Tests

Before testing the key hypotheses, we performed two preliminary tests. First, participants' facial averageness predicted facial similarity with the partner ($\gamma(SE) = 0.44(0.02)$, $p < .001$). We note that since all quantitative variables have been scaled, this effect size means that for every 1 standard deviation increase in facial averageness, facial similarity increases by 0.44 standard deviations. This relationship held even when accounting for ethnicity of the participant and ethnic concordance (i.e. both the participant and partner identifying as the same ethnicity) (γ(SE) = 0.44(0.02), $p < .001$). The same pattern of results held for sex-controlled facial similarity: facial averageness was a significant predictor (γ(SE) = 0.51(0.02), $p < .001$) and remained significant when controlling for ethnicity and ethnic concordance (γ(SE) = 0.51(0.02), $p < .001$). Therefore, both facial averageness and facial similarity (or sex-controlled similarity) were included in models together where relevant, so that their unique contributions could be estimated. The interaction term of facial averageness and similarity was not a significant predictor in any model and was therefore excluded from analyses.

3. 2. Facial Attractiveness

3. 2. 1. Facial Averageness and Similarity

When in separate models, facial averageness, and facial similarity each significantly predicted participants' rated facial attractiveness (γ(SE) = 0.09(0.03), p <.001, and γ(SE) = 0.06(0.02), $p = .016$, respectively). But when averageness and similarity were added into the same model, only averageness was a significant predictor of facial attractiveness (γ(SE) = 0.08(0.03), $p = .008$) and facial similarity was non-significant (γ(SE) = 0.03(0.02), $p = .186$). Averageness also became non-significant when we additionally controlled for the ethnicity of the rater and ethnic concordance (γ(SE) = 0.06(0.03), $p = .057$). This suggests that the variance shared between averageness and similarity, which is excluded when both variables are included as predictors in the same model, is driving the prediction of attractiveness. We confirmed this with a model fit test, which showed that including both facial averageness and similarity in in addition to the control variables (i.e. sex, ethnicity, and ethnic matching) explained more variance than the control variables alone (χ^2_2 = 6.83, $p = .033$). Overall, we can see that averageness and similarity together influence facial attractiveness ratings, but we cannot distinguish which of the variables is being used by participants to judge attractiveness. The same analyses were conducted for sex-controlled facial similarity instead of (raw) facial similarity. The pattern of results was the same (see Supplementary Table 24) – indeed, the magnitude of the effect sizes were slightly larger in general.

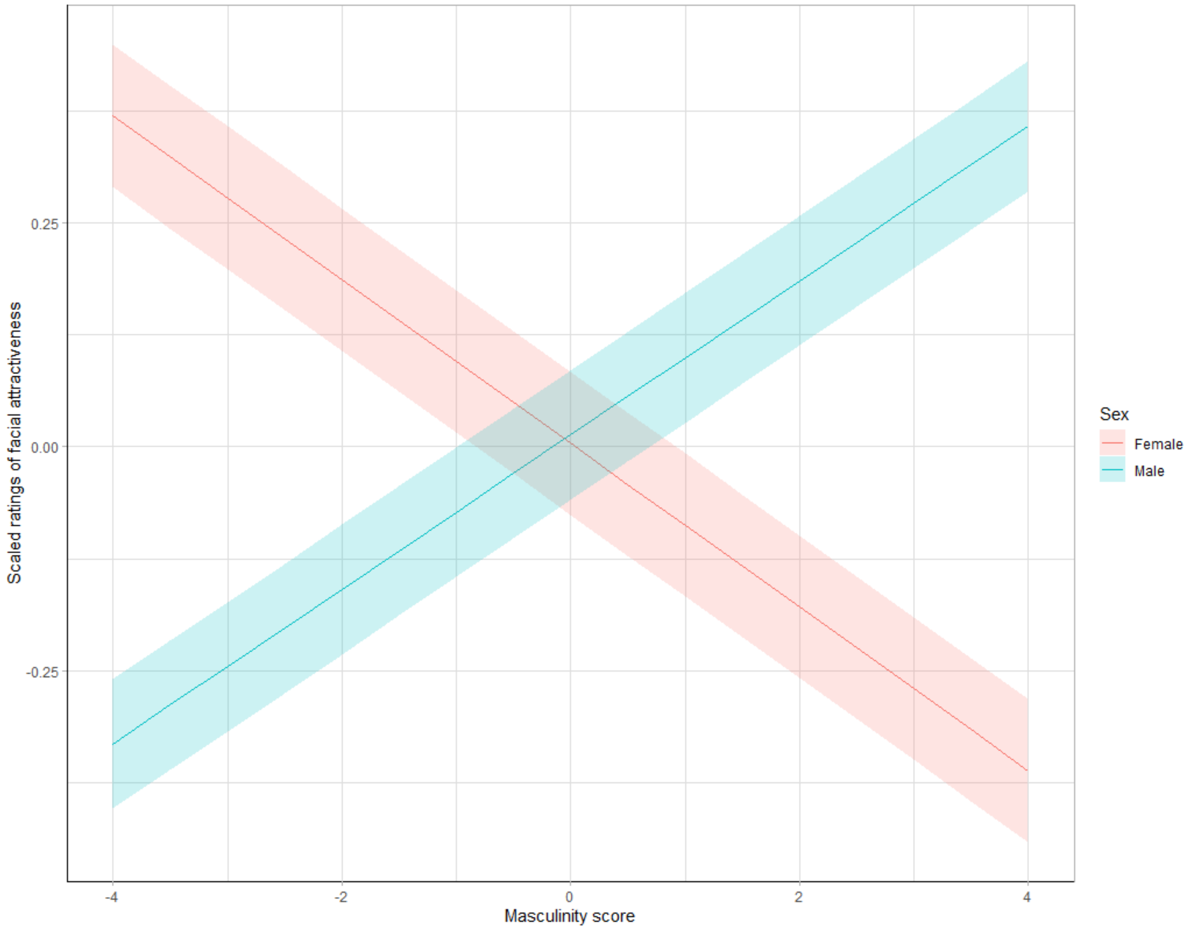
We also found participants received higher facial attractiveness ratings if they were rated by partners of the same ethnicity relative to being rated by a partner of a different ethnicity. This effect was larger when raw facial similarity was included in the model compared to the inclusion of sex-controlled similarity (γ(SE) = 0.040(0.02), p = .043, and γ(SE) = 0.036(0.02), p = .067, respectively).

3. 2. 2. Facial Masculinity

Overall, both with and without controlling for ethnicity and ethnic concordance, facial masculinity did not significantly predict facial attractiveness (γ(SE) = -0.01(0.04), p = .845, γ(SE) = 0.00(0.04), p = .985, respectively); but because this main effect averages over the expected different effects in each sex, it is the interaction between masculinity and sex that is more relevant, and this interaction was significant (γ(SE) = 0.11(0.04), p = .006, γ(SE) = 0.13(0.04), p = .002, respectively). Further analysis of the interaction effect revealed that in males, masculinity scores positively predict rated facial attractiveness, (γ(SE) = 0.09(0.04), p = .034), and in females, masculinity scores negatively predicted rated facial attractiveness (γ(SE) = -0.09(0.04), p = .014) (Figure 4). But when controlling for ethnicity and ethnic concordance, the previous significant association in males became non-significant (γ(SE) = 0.06(0.04), p = .109), although the negative association remained in females (γ(SE) = -0.08(0.04), p = .023). There was no evidence for (dis)assortative preferences for masculinity either with or without controlling for ethnicity and ethnic concordance (γ(SE) = -0.04(0.04), p = .292 and γ(SE) = -0.04(0.04), p = .292, respectively). Participants in ethnically concordant interactions also received higher facial attractiveness ratings compared to ethnically discordant interactions (γ(SE) = 0.04(0.02), p = .042).

Figure 4.

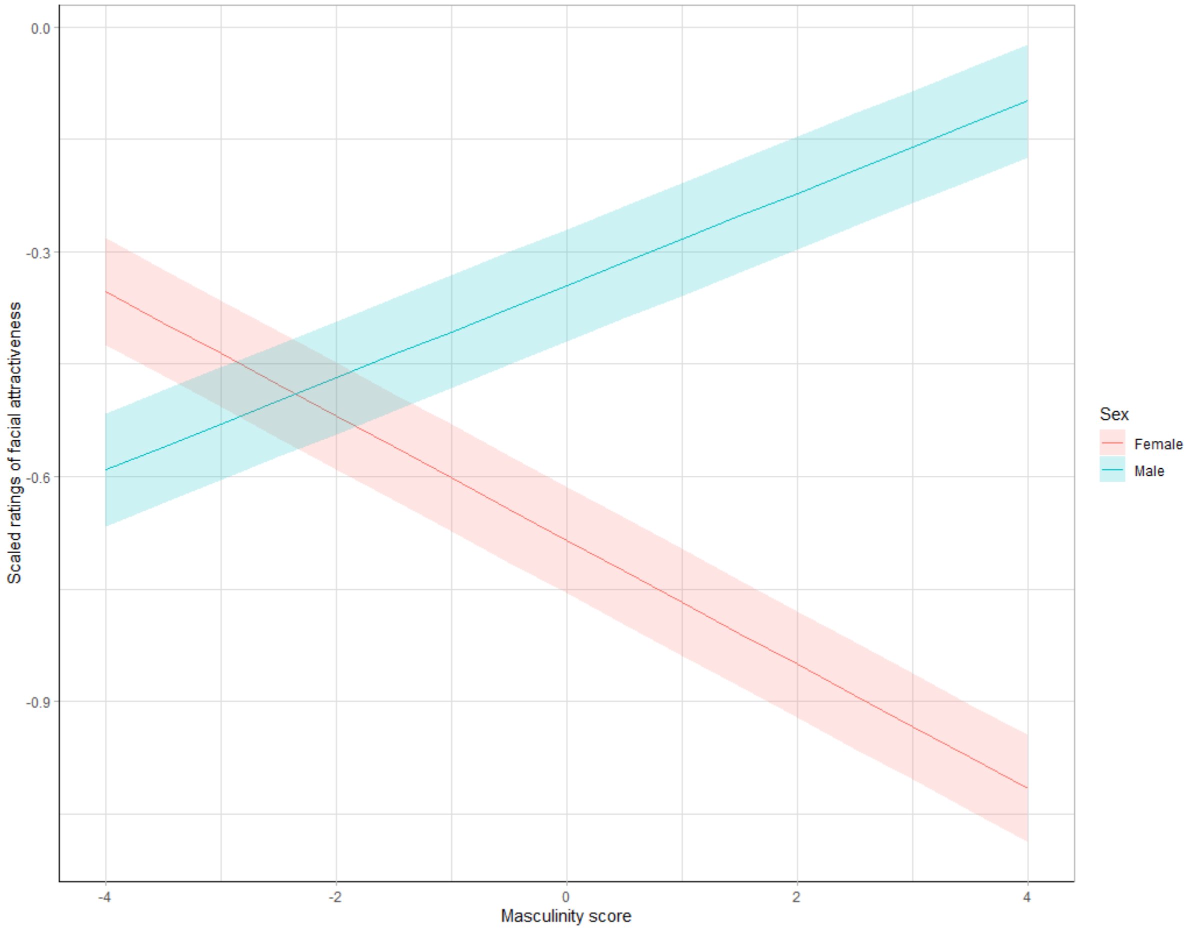
The interaction effect between sex and objective masculinity score on ratings of facial attractiveness. Coloured bands show 95% confidence intervals.



Note. The red and blue lines indicate the significant negative and positive association between masculinity and rated facial attractiveness in females and males, respectively.

Figure 5.

The interaction effect between sex and objective masculinity score on ratings of facial attractiveness when controlling for ethnicity and ethnic matching. Coloured bands show 95% confidence intervals.



Note. The red and blue lines indicate the significant negative and non-significant association between masculinity and rated facial attractiveness in females and males, respectively.

3. 3. Kindness and Understanding

Facial averageness did not predict ratings of kindness and understanding (γ(SE) = 0.02(0.02), p = .374), but facial similarity did (γ(SE) = 0.07(0.02), p <.001). The effects of facial similarity remained robust even when averageness, ethnicity, and ethnic concordance were in the same model (γ(SE) = 0.08(0.03), p = .002), and averageness remained non-significant (γ(SE) = -0.02(0.02), p = .453).

The same analyses were conducted for sex-controlled facial similarity instead of (raw) facial similarity. The pattern of results was the same (see Supplementary Table 23) – indeed, the magnitude of the effect sizes were slightly larger in general.

Ethnic concordance was predictive of kindness and understanding, though this effect was stronger when using raw similarity in the model compared to sex-controlled similarity (γ(SE) = 0.043(0.022), p = .046, and, γ(SE) = 0.036(0.022), p = .099, respectively).

4. Discussion

With a highly powered sample (n = 682), we show that during short face-to-face interactions, objective measures of facial traits – that may not be consciously salient or even detectable – predict evaluations of facial attractiveness and of kindness and understanding. These findings address major limitations in past studies which involved participants rating a series of photographs or computer-generated faces, as it was not clear that findings from those studies would generalise to real life interactions where people are moving, talking, changing facial expressions, displaying their personality, and so on. Additionally, our use of objective measures of averageness, similarity, and masculinity avoids biases and issues associated with participant ratings of those characteristics, which were often used in previous studies.

Though we found significant main effects of objectively measured facial traits on real-life ratings, we also observe that these effects are very small: the maximum standardised effect in the main analyses was 0.13, meaning that attractiveness increases 0.13 standard deviations for each standard deviation increase in the independent variable. Such small effect sizes are expected given the nature of our naturalistic study as there would be multiple sources of interference from extraneous variables. In addition, our measures are imperfect (we discuss limitations later); the associated noise also reduces the effect size estimates. Nonetheless, it is clear that attractiveness effects of specific facial metrics are subtle in the context of real-life interactions, in contrast to commonly large effects obtained in highly controlled studies using images.

In past facial attractiveness studies, the effects of facial averageness and facial similarity have been considered independently. If we consider them independently, we find that averageness and similarity each predict attractiveness ratings. (This finding is notable, given that similarity between raters and unmanipulated faces has not been studied as a contributor to attractiveness ratings.) But we also confirmed that objectively measured facial averageness and facial similarity would be substantially correlated, which is consistent with Penton-Voak et al. (1999)'s ideas. By definition, the average minimises the squared deviations from other individuals in a sample, and our results showed that the more average a face is, the more similar that face tends to be to other faces. Therefore, failure to account for this relationship may lead to one confounding the other's association with attractiveness. Indeed, when ethnicity and ethnic concordance were controlled for, neither facial averageness nor similarity individually predicted facial attractiveness, but when combined, averageness and similarity explained more variance in facial attractiveness than the control variables alone. Thus it was not possible to determine whether the significant combined effect was driven by attractiveness or similarity, or both. Our finding raises the possibility that previous findings that more average faces are rated more attractive (Cai et al., 2019; Foo et al., 2017; Lee et al., 2016; Light et al., 2016; Rhodes, 2006) might be in part (or even in whole) explained by the fact that more average faces tend to be more similar to the raters' faces.

DeBruine (2005) found that participants rated facial images as more trustworthy when the images were morphed to be more similar to the participant's own face; our findings were concordant, with participants rating more facially similar partners as more kind and understanding. This finding was robust and remained even when controlling for ethnicity and ethnic concordance, as well as averageness. Indeed, facial averageness, which is thought to indicate genetic quality, was not significantly associated with ratings of kindness and understanding, even without controlling for similarity. If humans do prefer similar faces in romantic partners, our findings support the idea that this preference may be driven by perceived prosociality (e.g. trustworthiness; kindness and understanding). According to theory, it makes sense that similarity predicts ratings of kindness and understanding while averageness does not; the information putatively revealed by averageness is relevant to mating but not to prosociality, whereas facial similarity may be relevant to perception of prosociality because of its overlap with kinship.

Past studies have generally been limited by the range of ethnicities assessed. Participants in DeBruine's (2005) study were diverse but limited to rating images of morphed faces that belonged to their own "phenotypic category" (ethnicity), whereas other studies have generally used only Caucasian participants/stimuli (e.g. Foo et al., 2017; Smith et al., 2006), limited their participants/stimuli to differ across two ethnicities (e.g. Perrett et al., 1998; G. Rhodes et al., 2001), or have not specified the ethnicities of the participants (e.g. Gillian Rhodes et al., 2001; Scott et al., 2010). By including a diverse range of ethnicities in the speed-dating interactions, our study was more representative of the demographics of Australia and other similar Western countries; on the other hand it posed analytic challenges given the pseudo-categorical nature of the ethnicity variable. We controlled for different ethnicities and ethnic concordance, but we were limited by the response range on the questionnaire (i.e. Caucasian, Asian, African, or Other: Please specify). Nevertheless, most participants who selected "Other" also specified their ethnicity as an open response. Based on these open responses, the first author further classified those participants into several main ethnic categories (see Supplementary materials for additional details). Having coded ethnicity in this manner, we found evidence of a revealed preference for ethnic concordance, such that participants who receive ratings from partners of the same ethnicity were rated higher in facial attractiveness, and kindness and understanding.

Consistent with past findings (Alharbi et al., 2020; Foo et al., 2017; Lee, Dubbs, et al., 2014; Rhodes, 2006), females with low facial masculinity scores (i.e. high femininity) were rated more facially attractive than females with high facial masculinity; this effect was robust and remained when controlling for ethnicity and ethnic concordance. In males, there was a positive association between masculinity and rated facial attractiveness, but this effect was no longer significant once we controlled for ethnicity and ethnic concordance, suggesting it may have been due to a confound involving ethnicity. Past studies have shown positive, negative, and null effects of masculinity on attractiveness in males (Foo et al., 2017; Lee, Mitchem, et al., 2014; Penton-Voak & Chen, 2004; Rhodes, 2006) – our null findings in real interactions may lend weight to a null or positive association given the direction of our estimate. However, we remain wary of overinterpreting the null effect of masculinity in men (after controlling for ethnicity and ethnic concordance) as compared to the significant positive effect in women, as the difference in the magnitude of these effects is only 0.02 despite the difference in statistical significance. We also found no evidence for assortative (or disassortative) preferences regarding facial masculinity in speed-dating dyads, consistent with findings by Burriss et al. (2011) who found that couples did not pair up (dis)assortatively with regard to facial masculinity.

Though our use of objective measures of face shape has major advantages in terms of eliminating biases that arise in ratings of facial characteristics, these measures also have limitations. For example, they exclude features such as eye, skin, and hair colour and texture, which are relevant to facial attractiveness. They also rely on 2D images; not only does this lose important facial information such as facial projection, but it also introduces potential biases if participants tilt their faces at different angles while being photographed (Sedgewick et al., 2017). These limitations could be improved upon by obtaining images of participants' side profiles or even 3D images that can subsequently be landmarked and/or rated.

Our current findings reflect preferences as revealed by ratings of potential partners, but it is unclear whether these preferences are realised when it comes to mate choice. Given that certain facial traits are preferred on average, we might expect individuals to pursue and partner up with those who possess these traits. But as realised mate choice involves the *mutual* desire to form a romantic partnership, these decisions are constrained by the mating market. For instance, individuals with high mate value are unlikely to pair with low mate value individuals. Given that not everyone can satisfy their preferences, individuals must make compromises. Therefore, it would be worthwhile to investigate how the objective facial characteristics we found to relate to attractiveness here may or may not also relate in real couples.

5. References

Abel, E. L., & Kruger, M. L. (2011). Facial resemblances between heterosexual, gay, and lesbian couples. *Psychol Rep*, *108*(3), 688-692. https://doi.org/10.2466/07.PR0.108.3.688-692

Adams, D. C., Collyer, M. L., Kaliontzopoulou, A., & Baken, E. K. (2022). *Geomorph: Software for geometric morphometric analyses. R package version 4.0.4*. In https://cran.r-project.org/package=geomorph

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2004). Geometric morphometrics: Ten years of progress following the 'revolution'. *Italian Journal of Zoology*, *71*(1), 5-16. https://doi.org/10.1080/11250000409356545

Alharbi, S. A. H., Holzleitner, I. J., Lee, A. J., Saribay, S. A., & Jones, B. C. (2020). Women's Preferences for Sexual Dimorphism in Faces: Data from a Sample of Arab Women. *Evolutionary Psychological Science*, *6*(4), 328-334. https://doi.org/10.1007/s40806-020-00244-y

Alvarez, L., & Jaffe, K. (2004). Narcissism Guides Mate Selection: Humans Mate Assortatively, as Revealed by Facial Resemblance, following an Algorithm of "Self Seeking Like". *Evolutionary Psychology*, *2*(1). https://doi.org/10.1177/147470490400200123

Baken, E. K., Collyer, M. L., Kaliontzopoulou, A., & Adams, D. C. (2021). geomorph v4.0 and gmShiny: enhanced analytics and a new graphical interface for a comprehensive morphometric experience. *Methods in Ecology and Evolution*, *12*, 2355-2363.

Bardua, C., Felice, R. N., Watanabe, A., Fabre, A. C., & Goswami, A. (2019). A Practical Guide to Sliding and Surface Semilandmarks in Morphometric Analyses. *Integr Org Biol*, *1*(1), obz016. https://doi.org/10.1093/iob/obz016

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-efects models using S4 classes*. In

Bittles, A. H., & Neel, J. V. (1994). The costs of human inbreeding and their implications for variations at the DNA level. *Nat Genet*, *8*(2), 117-121. https://doi.org/10.1038/ng1094-117

Burriss, R. P., Roberts, S. C., Welling, L. L., Puts, D. A., & Little, A. C. (2011). Heterosexual romantic couples mate assortatively for facial symmetry, but not masculinity. *Pers Soc Psychol Bull*, *37*(5), 601-613. https://doi.org/10.1177/0146167211399584

Cai, Z., Hahn, A. C., Zhang, W., Holzleitner, I. J., Lee, A. J., DeBruine, L. M., & Jones, B. C. (2019). No evidence that facial attractiveness, femininity, averageness, or coloration are cues to susceptibility to infectious illnesses in a university sample of young adult women. *Evolution and Human Behavior*, *40*(2), 156-159. https://doi.org/10.1016/j.evolhumbehav.2018.10.002

Cunningham, M. R., Roberts, A. R., Barbee, A. P., Druen, P. B., & Wu, C.-H. (1995). "Their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology*, *68*(2), 261-279. https://doi.org/10.1037/0022-3514.68.2.261

DeBruine, L. (2016). *Young adult composite faces*. figshare. https://figshare.com/articles/figure/Young_adult_composite_faces/4055130

DeBruine, L. (2018). *Webmorph: Beta Release 2*. In (Version v0.0.0.9001) Zenodo.

DeBruine, L. M. (2002). Facial resemblance enhances trust. *Proc Biol Sci*, *269*(1498), 1307-1312. https://doi.org/10.1098/rspb.2002.2034

DeBruine, L. M. (2004). Facial resemblance increases the attractiveness of same-sex faces more than other-sex faces. *Proc Biol Sci*, *271*(1552), 2085-2090. https://doi.org/10.1098/rspb.2004.2824

DeBruine, L. M. (2005). Trustworthy but not lust-worthy: context-specific effects of facial resemblance. *Proc Biol Sci*, *272*(1566), 919-922. https://doi.org/10.1098/rspb.2004.3003

Dobzhansky, T. (1982). *Genetics and the origin of species*.

Everett, J. A., Faber, N. S., & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. *Front Behav Neurosci*, *9*, 15. https://doi.org/10.3389/fnbeh.2015.00015

Foo, Y. Z., Simmons, L. W., & Rhodes, G. (2017). Predictors of facial attractiveness and health in humans. *Sci Rep*, *7*, 39731. https://doi.org/10.1038/srep39731

Geldart, S., Maurer, D., & Carney, K. (1999). Effects of eye size on adults' aesthetic ratings of faces and 5-month-olds' looking times. *Perception*, *28*(3), 361-374. https://doi.org/10.1068/p2885

Griffiths, R. W., & Kunz, P. R. (1973). Assortative mating: a study of physiognomic homogamy. *Soc Biol*, *20*(4), 448-453. https://doi.org/10.1080/19485565.1973.9988075

Hamilton, W. D. (1964). The genetical evolution of social behaviour. I. *J Theor Biol*, *7*(1), 1-16. https://doi.org/10.1016/0022-5193(64)90038-4

Hansen, F., DeBruine, L. M., Holzleitner, I. J., Lee, A. J., O'Shea, K. J., & Fasolt, V. (2020). Kin recognition and perceived facial similarity. *J Vis*, *20*(6), 18. https://doi.org/10.1167/jov.20.6.18

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning* (2 ed.). Springer New York, NY. https://doi.org/https://doi.org/10.1007/978-0-387-84858-7

Helgason, A., Palsson, S., Gudbjartsson, D. F., Kristjansson, T., & Stefansson, K. (2008). An association between the kinship and fertility of human couples. *Science*, *319*(5864), 813-816. https://doi.org/10.1126/science.1150232

Hill, A. K., Hunt, J., Welling, L. L. M., Cárdenas, R. A., Rotella, M. A., Wheatley, J. R., Dawood, K., Shriver, M. D., & Puts, D. A. (2013). Quantifying the strength and form of sexual selection on men's traits. *Evolution and Human Behavior*, *34*(5), 334-341. https://doi.org/10.1016/j.evolhumbehav.2013.05.004

Hinsz, V. B. (2016). Facial Resemblance in Engaged and Married Couples. *Journal of Social and Personal Relationships*, *6*(2), 223-229. https://doi.org/10.1177/026540758900600205

Holzleitner, I. J., & DeBruine, L. M. (2020). *facefuns: Convenience Functions Around 2D and 3D Face Shape Analysis*. In https://github.com/iholzleitner/facefuns

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2020). *lmerTest: Tests in Linear Mixed Effects Models*. In

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1). https://doi.org/10.2307/2529310

Langlois, J. H., Ritter, J. M., Roggman, L. A., & Vaughn, L. S. (1991). Facial diversity and infant preferences for attractive faces. *Developmental Psychology*, *27*(1), 79-84. https://doi.org/10.1037/0012-1649.27.1.79

Langlois, J. H., Roggman, L. A., Casey, R. J., Ritter, J. M., Rieser-Danner, L. A., & Jenkins, V. Y. (1987). Infant preferences for attractive faces: Rudiments of a stereotype? *Developmental Psychology*, *23*(3), 363-369. https://doi.org/10.1037/0012-1649.23.3.363

Lee, A. J., Dubbs, S. L., Von Hippel, W., Brooks, R. C., & Zietsch, B. P. (2014). A multivariate approach to human mate preferences. *Evolution and Human Behavior*, *35*(3), 193-203. https://doi.org/10.1016/j.evolhumbehav.2014.01.003

Lee, A. J., Mitchem, D. G., Wright, M. J., Martin, N. G., Keller, M. C., & Zietsch, B. P. (2014). Genetic factors that increase male facial masculinity decrease facial

attractiveness of female relatives. *Psychol Sci*, *25*(2), 476-484. https://doi.org/10.1177/0956797613510724
Lee, A. J., Mitchem, D. G., Wright, M. J., Martin, N. G., Keller, M. C., & Zietsch, B. P. (2016). Facial averageness and genetic quality: Testing heritability, genetic correlation with attractiveness, and the paternal age effect. *Evol Hum Behav, 37*(1), 61-66. https://doi.org/10.1016/j.evolhumbehav.2015.08.003
Light, L. L., Hollander, S., & Kayra-Stuart, F. (2016). Why Attractive People are Harder to Remember. *Personality and Social Psychology Bulletin*, *7*(2), 269-276. https://doi.org/10.1177/014616728172014
Lindova, J., Little, A. C., Havlicek, J., Roberts, S. C., Rubesova, A., & Flegr, J. (2016). Effect of Partnership Status on Preferences for Facial Self-Resemblance. *Front Psychol*, *7*, 869. https://doi.org/10.3389/fpsyg.2016.00869
Penton-Voak, I. S., & Chen, J. Y. (2004). High salivary testosterone is linked to masculine male facial appearance in humans. *Evolution and Human Behavior*, *25*(4), 229-241. https://doi.org/10.1016/j.evolhumbehav.2004.04.003
Penton-Voak, I. S., Perrett, D. I., & Peirce, J. W. (1999). Computer graphic studies of the role of facial similarity in judgements of attractiveness. *Current Psychology*, *18*(1), 104-117. <Go to ISI>://WOS:000080926000009
Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., Henzi, S. P., Castles, D. L., & Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature*, *394*(6696), 884-887. https://doi.org/10.1038/29772
Perrett, D. I., May, K. A., & Yoshikawa, S. (1994). Facial shape and judgements of female attractiveness. *Nature, 368*(6468), 239-242. https://doi.org/10.1038/368239a0
R Development Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. In www.R-project.org
Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annu Rev Psychol, 57*, 199-226. https://doi.org/10.1146/annurev.psych.57.102904.190208
Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., & Akamatsu, S. (2001). Attractiveness of facial averageness and symmetry in non-western cultures: in search of biologically based standards of beauty. *Perception*, *30*(5), 611-625. https://doi.org/10.1068/p3123
Rhodes, G., Zebrowitz, L. A., Clark, A., Kalick, S. M., Hightower, A., & McKay, R. (2001). Do facial averageness and symmetry signal health? *Evolution and Human Behavior*, *22*(1), 31-46. https://doi.org/10.1016/s1090-5138(00)00060-x
Rubenstein, A. J., Kalakanis, L., & Langlois, J. H. (1999). Infant preferences for attractive faces: a cognitive explanation. *Dev Psychol*, *35*(3), 848-855. https://doi.org/10.1037//0012-1649.35.3.848
Scott, I. M., Pound, N., Stephen, I. D., Clark, A. P., & Penton-Voak, I. S. (2010). Does masculinity matter? The contribution of masculine face shape to male attractiveness in humans. *PLoS One*, *5*(10), e13585. https://doi.org/10.1371/journal.pone.0013585
Sedgewick, J. R., Flath, M. E., & Elias, L. J. (2017). Presenting Your Best Self(ie): The Influence of Gender on Vertical Orientation of Selfies on Tinder. *Front Psychol*, *8*, 604. https://doi.org/10.3389/fpsyg.2017.00604
Sidari, M. J., Lee, A. J., Murphy, S. C., Sherlock, J. M., Dixson, B. J. W., & Zietsch, B. P. (2020). Preferences for Sexually Dimorphic Body Characteristics Revealed in a Large Sample of Speed Daters. *Social Psychological and Personality Science*, *12*(2), 225-236. https://doi.org/10.1177/1948550619882925
Smith, M. J., Perrett, D. I., Jones, B. C., Cornwell, R. E., Moore, F. R., Feinberg, D. R., Boothroyd, L. G., Durrani, S. J., Stirrat, M. R., Whiten, S., Pitman, R. M., & Hillier,

S. G. (2006). Facial appearance is a cue to oestrogen levels in women. *Proc Biol Sci*, *273*(1583), 135-140. https://doi.org/10.1098/rspb.2005.3296

Tajfel, H. (1970). Experiments in Intergroup Discrimination. *Scientific American*, *223*(5), 96-103. https://doi.org/DOI 10.1038/scientificamerican1170-96

Tea-Makorn, P. P., & Kosinski, M. (2020). Spouses' faces are similar but do not become more similar with time. *Sci Rep*, *10*(1), 17001. https://doi.org/10.1038/s41598-020-73971-8

Valentine, K. A., Li, N. P., Penke, L., & Perrett, D. I. (2014). Judging a Man by the Width of His Face The Role of Facial Ratios and Dominance in Mate Choice at Speed-Dating Events. *Psychological Science*, *25*(3), 806-811. https://doi.org/10.1177/0956797613511823

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Weller, M., & Santos, S. (2013). A positive association between consanguinity and fertility in communities of Paraiba, Northeast Brazil. *Ann Hum Biol*, *40*(6), 527-530. https://doi.org/10.3109/03014460.2013.815271

Zajonc, R. B., Adelmann, P. K., Murphy, S. T., & Niedenthal, P. M. (1987). Convergence in the physical appearance of spouses. *Motivation and Emotion*, *11*(4), 335-346. https://doi.org/10.1007/bf00992848

Zelditch, M. L., Swiderski, D. L., Sheets, H. D., & Fink, W. L. (2004). *Geometric Morphometrics for Biologists*. https://doi.org/10.1016/b978-0-12-778460-1.X5000-5

Zietsch, B. P., Verweij, K. J., Heath, A. C., & Martin, N. G. (2011). Variation in human mate choice: simultaneously investigating heritability, parental influence, sexual imprinting, and assortative mating. *Am Nat*, *177*(5), 605-616. https://doi.org/10.1086/659629