Social evaluation of faces across gender and familiarity

Mila Mileva, Robin S. S. Kramer[1] & A. Mike Burton

Department of Psychology, University of York, UK

Correspondence to:

Mila Mileva

Department of Psychology

University of York

Heslington, York

YO10 5DD, UK

mila.mileva@york.ac.uk

Running head: Social evaluation across gender and familiarity

1. Present address: School of Psychology, University of Lincoln, UK

Abstract

Models of social evaluation aim to capture the information people use to form first impressions of unfamiliar others. However, little is currently known about the relationship between perceived traits across gender. In Study 1, we asked viewers to provide ratings of key social dimensions (dominance, trustworthiness etc.) for multiple images of 40 unfamiliar identities. We observed clear sex differences in the perception of dominance – with negative evaluations of high dominance in unfamiliar females, but not males. In Study 2, we used the social evaluation context to investigate key predictions about the importance of pictorial information in familiar and unfamiliar face processing. We compared the consistency of ratings attributed to different images of the same identities and demonstrated that ratings of images depicting the same familiar identity are more tightly clustered than those of unfamiliar identities. Such results imply a shift from *image* rating to *person* rating with increased familiarity, a finding which generalises results previously observed in studies of identification.

Keywords: Social evaluation; first impressions; face perception; familiar faces

Introduction

Although we are often reminded not to judge a book by its cover, people have been shown to form stable first impressions from faces within a few milliseconds (Willis & Todorov, 2006). These evaluations affect our choices and behaviours not only in situations where appearance might be relevant, but also in situations as diverse as criminal sentencing (Eberhardt et al., 2006; Wilson & Rule, 2016), political elections (Olivola & Todorov, 2010) and employment and finance decisions (Graham, Harvey, & Puri, 2016; Rule & Ambady, 2008; 2009). While the accuracy of these personality evaluations is generally limited (Todorov et al., 2015; although see Kramer & Ward, 2010), viewers are consistent in their judgements, implying that they are using some physical information in the face to inform their impressions (Kramer, Mileva, & Ritchie, 2018; Zebrowitz & Montepare, 2008). Oosterhof and Todorov (2008) collected ratings of systematically varied computer-graphics faces on a range of social traits and identified two underlying dimensions in social evaluation: trustworthiness and dominance. Recent studies making use of more natural, 'ambient' images, varying in emotional expression, pose, lighting and camera angle observed the same trustworthiness and dominance dimensions, as well as an additional third – youthful-attractiveness (Sutherland et al., 2013). While these models appear to have captured the fundamental face evaluation components and they fit well with other social evaluation

models such as concept evaluation (Osgood, Suci, & Tannenbaum, 1957) and interpersonal perception (Wiggins, 1979), little is known about the relationship between these traits across gender and familiarity.

Gender differences in social evaluation

The importance of gender for social evaluation is highlighted in studies from the social stereotypes literature (Imhoff et al., 2013; Oldmeadow, Sutherland, & Young, 2013) building upon the process of categorisation (Secord, 1959). In the context of face evaluation, first impressions are the product of assigning a category to a specific face and using category-associated information to form one's social judgements. Given the similarities between face and general social evaluation models, it is possible that the fundamental social dimensions – trustworthiness and dominance – are attributed differently for male and female faces. This is supported by Sutherland, Young, et al. (2015), who collected ratings of male and female face images with stereotypical (e.g. female images rated high on the femininity scale) and counter-stereotypical (e.g. female images rated high on the masculinity scale) appearance. On the one hand, masculine-looking female faces were perceived as significantly more dominant, less attractive and less trustworthy than feminine-looking female faces. This implies a positive relationship between attractiveness and trustworthiness, and a negative relationship between each of

those traits and dominance for female faces. On the other hand, masculine-looking male faces received significantly higher dominance ratings and significantly lower attractiveness ratings. However, there was no difference in trustworthiness attribution, demonstrating a different pattern of results for male identities. These findings were further supported with a second study where masculinity/femininity was not manipulated, yet female faces high in dominance were rated as significantly less trustworthy than female faces low in dominance, whereas no such difference was found between high- and low-dominance male images.

Social evaluation and familiarity

By its nature, first impressions research focuses on the perception of faces that are unfamiliar to the viewer. However, influential models of familiar face recognition (Bruce & Young, 1986; Burton, Bruce, & Hancock, 1999) postulate that recognition of a known person automatically activates cross-modal and semantic information such as that person's voice, name, occupation and so on. This proposal is supported by evidence from a variety of priming studies, typically showing enhanced access to personal characteristics as a result of prior, but not overlapping, exposure to a known person (e.g., Burton, Kelly, & Bruce, 1998; Schweinberger, Pfutze, & Sommer, 1995; Young, Hellawell, & De Haan, 1988). Judging the social characteristics of such a familiar

person, would inevitably incorporate all of this perceptual and semantic information which might lead to social ratings that reflect one's experience of the person, rather than ratings purely based on the physical characteristics of their face.

Nevertheless, recent work has highlighted the very large variability within different photos of the same face (Andrews, Jenkins, Cursiter, & Burton, 2015; Burton, 2013; Jenkins, White, Van Montfort, & Burton, 2011; Todorov & Porter, 2014). People are well-aware that certain photos make them look more or less attractive (for example) and use this knowledge to select different photos for different purposes (e.g. for work sites, dating sites etc.; Todorov & Porter, 2014; White, Sutherland, & Burton, 2017). We might therefore expect different photos of the same face, albeit showing a familiar person, to elicit somewhat varying social attributions. However, we hypothesise that there will be less variance in these social judgements to images of a familiar person, than the corresponding judgements to an unfamiliar person as they will be based on a combination of visual and semantic cues rather than visual cues only.

There is now considerable evidence suggesting that familiar and unfamiliar faces are processed differently. The bulk of this evidence relates to identification tasks, such as memory for faces and face matching (Bruce et al., 1999; Burton et al., 1999), and

typically indicates that unfamiliar faces are processed in a more image-bound way, whereas perception of familiar faces relies on more abstractive codes (Hancock et al., 2000; Johnston & Edmonds, 2009). It remains unclear whether familiar and unfamiliar faces are perceived in a qualitatively or quantitatively different way, and researchers have continued to examine the range of behavioural tasks over which this distinction manifests (e.g., Baker, Laurence, & Mondloch, 2017; Balas & Pearson, 2017; Laurence & Mondloch, 2016; Megreya & Burton, 2006). In this paper, we elicit social judgements both to familiar and unfamiliar faces, and examine the differences between these. Any observed differences will contribute to contemporary attempts to understand familiarity by extending its markers beyond identity-based tasks.

First impressions from average and composite faces

As we noted above, different images of the same unfamiliar person can give rise to quite widely varying social attributions (Todorov & Porter, 2014). Nevertheless, we can ask whether there is some central, core or modal attribution common to all images of that person. One way to examine this is to use graphical face-averaging, in which images of the same identity are morphed together (Burton, Jenkins, Hancock, & White, 2005; Jenkins & Burton, 2011). There is evidence that people extract an average-like representation when presented with sets of familiar or unfamiliar faces (Kramer,

Ritchie, & Burton, 2015; Neumann, Schweinberger, & Burton, 2013) and some advantages of face averages have been shown in both human and computer identity recognition (Jenkins & Burton, 2008; Ritchie et al., 2018; White et al., 2014; although see Ritchie, Mireku, & Kramer, in press).

The appeal of the averaging process comes from the fact that it eliminates superficial image information while preserving the core identity-diagnostic information as more images are incorporated into the average. This makes the average image a more stable identity representation (Jenkins & Burton, 2011). White et al. (2014) report that matching an average and an exemplar improves recognition compared to matching two exemplars and propose average images as an alternative form of photo ID. However, averaging has also been associated with certain artefacts such as blurring and smoothing of face texture, making average images qualitatively different from exemplars. What is more, this soft-focus effect, which removes any temporary skin imperfections, could influence social evaluation in terms of attractiveness, trustworthiness and even distinctiveness.

Some evidence for the possible effect of averages on social evaluation comes from studies using composite faces created by digitally blending many images together.

(Langlois & Roggman, 1990; Langlois, Roggman, & Musselman, 1994). A consistent finding is that composite images are judged as significantly more attractive than the individual exemplar images. Little and Hancock (2002), for example, used exemplar images of male identities and compared their attractiveness, distinctiveness and masculinity with facial composites. Their results showed that composites were perceived as more attractive as well as less distinctive and less masculine than the original images. Therefore, it is possible that averages may be socially evaluated in a qualitatively different way than normal exemplar images.

This is important because of the substantial number of studies demonstrating the impact of first impressions on people's decisions and behaviours, for example political elections (Ballew & Todorov, 2007; Sussman, Petkova, & Todorov, 2013; see Olivola & Todorov, 2010 for a review), criminal sentencing (Dumas & Teste, 2006), eyewitness testimony (Flowe & Humphries, 2011) and punishment severity (Wilson & Rule, 2016; see Todorov et al., 2015 for a review). If average images are used for official identification, as suggested by White et al. (2014), and they could be evaluated somewhat differently than exemplars, then it is possible that this will affect the decisions of those making the identification.

Study 1: Unfamiliar faces – gender differences and the effect of averaging

In the first study, we aim to explore social evaluation across gender, using unfamiliar faces only. Evidence from the social stereotypes literature highlights differences in the way males and females are socially evaluated, yet most influential face evaluation models (Oosterhof & Todorov, 2008; Walker & Vetter, 2009) use images of male and female identities together, leaving undetected any possible differences in ratings due to gender. Based on findings from Sutherland, Young, et al. (2015), we expect to see clear gender differences in the attribution of the fundamental social dimensions – trustworthiness and dominance. More specifically, we anticipate dominant female faces to be evaluated less favourably than dominant male faces. The study further aims to establish the social information conveyed by average images and how this information compares with ratings of normal exemplar images of the same identity. This is particularly relevant in light of recent studies suggesting average images might be a better alternative to photographic ID documents (White et al., 2014).

We collected ratings of attractiveness, trustworthiness, dominance, extraversion and distinctiveness for different images of the same unfamiliar identities (four exemplar images and one average image for each of 40 identities). Based on findings from the social stereotypes literature (Sutherland, Young, et al., 2015), we expect to identify

gender differences in the attribution of first impressions, specifically in the relationship between trustworthiness and dominance. We also explore the effect of averaging images of the same person together in order to establish whether a physical identity average would give rise to similar social ratings to those elicited by different exemplar images of the same identity. In short: do ratings of someone's average image correspond to average ratings for their individual images?

Method

Participants

A total of 27 participants (three male, $M = 21.6$ years, age range: 18-30) from the University of Aberdeen took part in the study. All participants had normal or corrected-to-normal vision and received payment or course credit for their participation. Informed consent was provided prior to their participation in accordance with the ethical standards stated in the 1964 Declaration of Helsinki. Experimental procedures were also approved by the psychology department ethics committee at the University of Aberdeen.

Materials

A set of 200 face images was used as experimental stimuli. This included four different exemplar images of 40 unfamiliar identities (20 male and 20 female) as well as an average of those four images for each identity (details of average construction to follow). The images depicted local celebrities from foreign countries, not known by British participants. Exemplar images were downloaded from a Google Images search by entering the name of the person and choosing the first four images that were in full colour, broadly frontal and with no parts of the face obscured by clothing or glasses. They were all naturally occurring or "ambient" images and captured a good amount of face variability due to differences in lighting, pose and emotional expressions. Images were cropped and resized to 380 x 570 pixels.

To construct average images, face shape was captured by manually indicating the *xy*-coordinates of 82 anatomical landmarks (e.g. inner corner of the eyes, centre of lower lip). These landmarked images were then co-registered by morphing the four images of one identity to a standard face template using bi-cubic interpolation. The average face texture was derived from the mean RGB values for each pixel and the average shape was derived from the mean *xy*-coordinates of each facial landmark. The average image for each identity was then created by morphing the average texture to the average shape for that corresponding identity (see Burton et al., 2005, for further details). Images were

morphed using custom MATLAB software (InterFace - Kramer, Jenkins, & Burton, 2016). See Figure 1 for exemplar and average image examples.

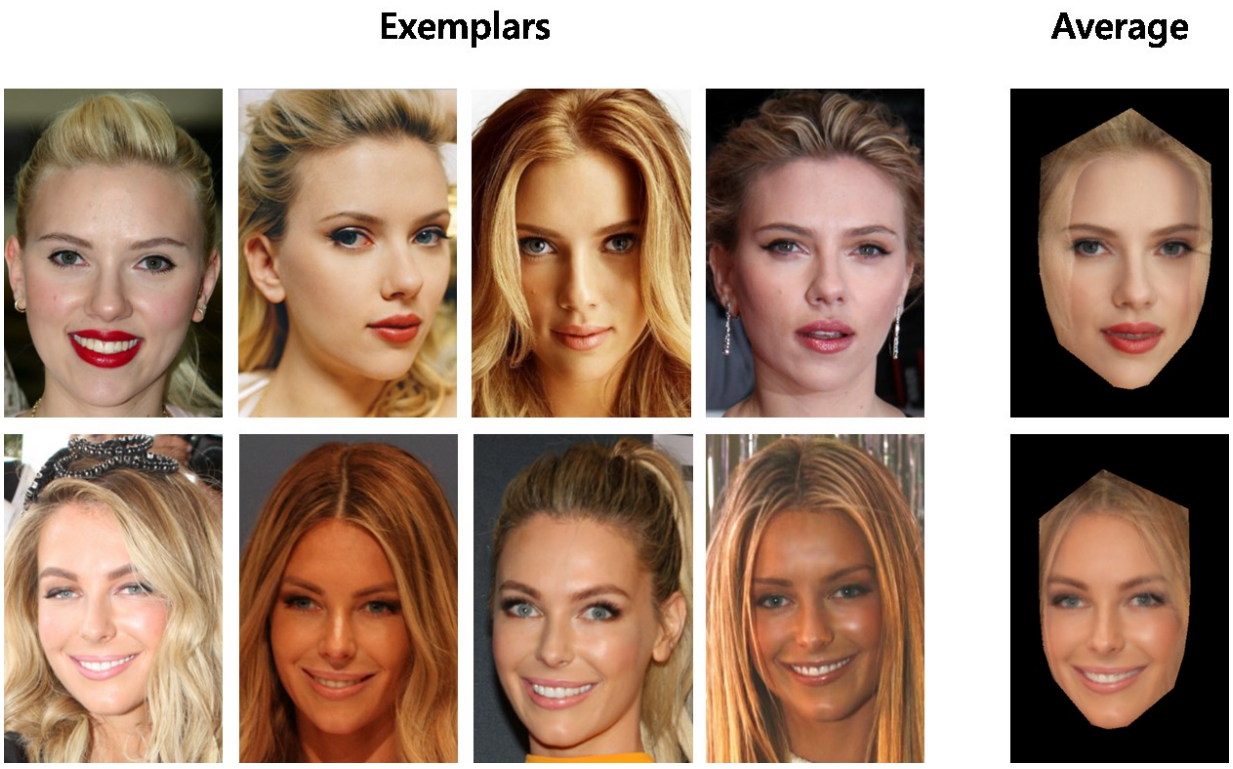

Figure 1. Exemplar and average images for a familiar (top row) and unfamiliar (bottom row) person. See image attributions in the Acknowledgements section.

Design and Procedure

Participants were tested individually on a standard PC running MATLAB R2014a, and stimuli were displayed on an 18-inch monitor. Participants were asked to rate all 200 images from the set. Ratings were collected for the three fundamental social evaluation dimensions - trustworthiness, dominance, and attractiveness. Images were also evaluated on distinctiveness in order to address predictions based on Little and Hancock (2002), as well as extraversion, which is part of the Big Five personality traits model (McCrae & Costa, 1987). Extraversion was used as a control measure since traits included in the Big Five, and extraversion particularly, show small to zero gender differences (see Lippa, 2010, for a review; Sutherland, Rowley, et al., 2015). All traits were rated on a 9-point scale where 1 represented the lower (e.g. not at all trustworthy) and 9 represented the higher (e.g. extremely trustworthy) end of the scale. Each face was presented individually at the centre of the screen with the rating scales positioned below the image, and participants rated the face for all attributes at their own pace before proceeding to the next face. Image presentation order was randomised individually for each participant.

Results and Discussion

Inter-trait Correlations

Table 1 shows Pearson's correlations between social traits attributed to unfamiliar male and female faces. The Holm-Bonferroni correction (1979) was applied to account for multiple comparisons. Correlations for male and female faces followed the same general pattern with a few key differences. While attractiveness and trustworthiness were positively correlated for both male and female faces, this relationship was significantly stronger for male faces ($z = 3.11$, $p = .001$). Figure 2A shows that this is due to more extended use of the scales for male faces: participants rate more males unattractive than they do females.

The other key gender difference we observed involved the relationship between dominance and trustworthiness (Table 1, Figure 2B). For female faces, dominance correlated negatively with trustworthiness whereas no such correlation was found for male faces. Gender differences in this relationship were close to significance ($z = 1.94$, $p = .052$). This demonstrates that female faces perceived as more dominant are also perceived as less trustworthy, whereas the perception of dominance in male faces does not seem to be related to the perception of their trustworthiness. Such findings fit well with our predictions, as well as previous literature (Sutherland, Young, et al., 2015), and suggest that male and female faces are evaluated somewhat differently on the fundamental social dimensions: trustworthiness and dominance. As predicted, no significant gender differences were found for extraversion and distinctiveness, which

act as controls to ensure that image differences do not give rise to systematic sex differences on any rating scale.

Table 1

Correlations Between Social Traits for Unfamiliar Male and Female Identities.

	Male Identities				Female Identities			
	A	T	Dom	E	A	T	Dom	E
Attractiveness	–				–			
Trustworthiness	**.76****	–			**.50****	–		
Dominance	.19	**-.17**	–		.26*	**-.43****	–	
Extraversion	.31*	.42**	.25	–	.46**	.41**	.18	–
Distinctiveness	.41**	.15	.21	.07	.49**	.14	.42**	.22

N = 100, * $p < .05$, ** $p < .001$ (Holm-Bonferroni corrected). Bold, underlined correlations are reliably different between male and female identities.

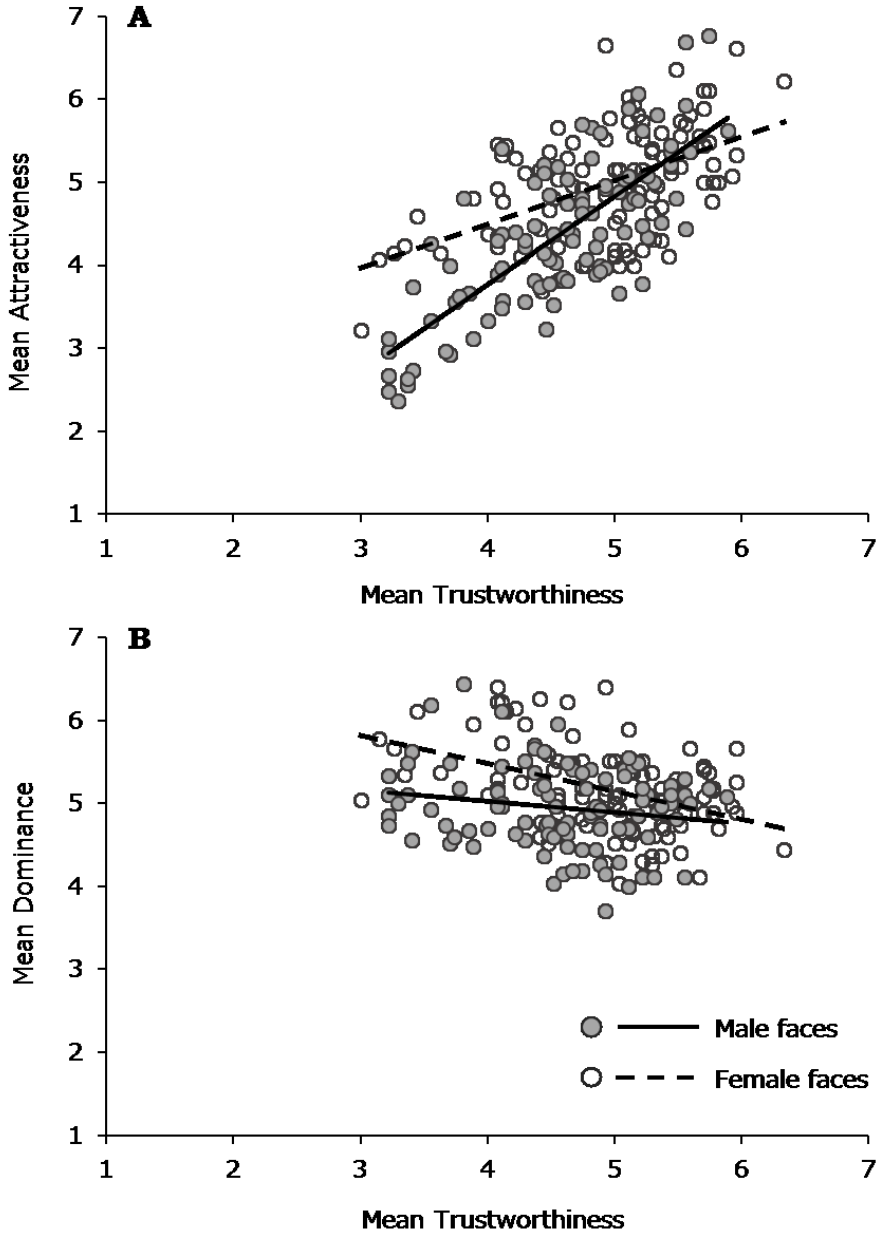

Figure 2. Correlations between attractiveness and trustworthiness (A) and between dominance and trustworthiness (B) across face gender.

First Impressions from Averages

Table 2 shows Pearson's correlations between ratings attributed to the average image of each person and the mean rating of that person's exemplar images. Additionally, we correlated the ratings attributed to the average images of each person with a randomly-selected exemplar of the same person. This procedure was carried out 1000 times with a different combination of exemplar images at each iteration. Table 2 also shows the mean correlation (across those 1000 runs, after an *r*-to-*z* transform) for each social trait, separately for male and female identities. Both analyses present us with a similar pattern of results - very high positive correlations for all attributes, which implies that overall, the physical average image corresponds to a social average. Such findings suggests that image averaging, while introducing artefacts such as smoothing and blurring (see Figure 1), nevertheless preserves the relative rank of faces across socially-important dimensions.

Table 2

Correlations between Ratings Attributed to the Physical Average and the Mean of the Four Exemplars as well as a Randomly-selected Exemplar.

	Male Faces		Female Faces	
	Mean of All Exemplars	Random Exemplar	Mean of All Exemplars	Random Exemplar
Attractiveness	.95	.83	.93	.66
Trustworthiness	.91	.75	.92	.77
Distinctiveness	.88	.77	.91	.64
Extraversion	.92	.81	.88	.64
Dominance	.92	.73	.80	.68

N = 20, all p_{max} = .002

Study 2: Familiar vs unfamiliar faces

In Study 2, we compare social evaluations of familiar and unfamiliar faces. We aim to establish whether the large familiarity effect observed in tests of identity is also present in social judgements. Specifically, if familiar and unfamiliar faces are processed in a

qualitatively different way, then we would expect different images of the same unfamiliar person to receive widely varying social ratings as they will be based on physical properties of the image. By contrast, images of the same familiar identity might elicit more consistent ratings, despite the superficial differences in images, as they will reflect a more abstractive representation of the *person* rather than the *face* which will include semantic as well as visual information.

Method

Participants

A total of 27 participants (5 male, $M = 20.2$ years, age range: 18-31) from the University of York took part in the study. All participants had normal or corrected-to-normal vision and received payment or course credit for their participation. Informed consent was provided prior to their participation in accordance with the ethical standards stated in the 1964 Declaration of Helsinki. Experimental procedures were also approved by the psychology department ethics committee at the University of York.

Materials

In addition to the unfamiliar face set from Study 1, we also collected another 200 familiar face images. This included four different exemplar images of 40 unfamiliar identities (20 male and 20 female) depicting well-known Hollywood celebrities and an average constructed for each identity (see Figure 1 for an example). The method by which images were collected and processed was the same as with the unfamiliar set in Study 1.

Design and Procedure

Participants were tested individually in labs at the University of York, equipped with a standard PC running MATLAB R2014a. The experiment followed the same procedure as in Study 1.

Analysis Strategy

Integrating rating data for both the familiar and unfamiliar face sets together, we created a five-dimensional social attribute space with dimensions corresponding to ratings of attractiveness, trustworthiness, dominance, distinctiveness and extraversion. In this space, if images of the same person are rated consistently, they will lie closer together, whereas images of the same person with very different social evaluations will be located further away from one another. In order to quantify the correspondence between social

ratings attributed to images of the same person, we used Procrustes analysis (Gower, 1975), separately for familiar and unfamiliar identities.

Procrustes analysis transforms the sets of social attributes in order to achieve maximal superimposition by minimising the sum of squares distances between the corresponding points in each set. The significance of the goodness-of-fit statistic is determined using a PROcrustean randomisation TEST (PROTEST; Jackson, 1995; Peres-Neto & Jackson, 2001), which estimates the probability of observing a given correspondence in comparison with a large number of equivalent values generated by randomly shuffling the original data set.

This procedure was carried out for 10,000 iterations where, for each iteration, two exemplar images were randomly selected for each identity. The goodness-of-fit for the two sets of social attribute ratings was measured and the 'by chance' equivalent for the two sets (i.e., the fit that is to be expected by chance) was produced by shuffling the attribute ratings and recalculating the goodness-of-fit. We used two different shuffling approaches – for the first, the location values within each trait were shuffled (Jackson, 1995), and for the second, the identity labels were shuffled (Peres-Neto & Jackson, 2001). Therefore, the observed goodness-of-fit and the two 'by chance' measures were calculated for each iteration. Moreover, in order to control for any potential problems

with scaling from the Procrustes analysis, we computed exactly the same transformation without the scaling component.

For additional consistency between Studies 1 and 2, we also analysed the inter-trait correlations separately for male and female familiar identities as well as the correlations between ratings attributed to the average and exemplar images. The results of these analyses should be interpreted with caution because the social evaluation literature is mostly based on zero-acquaintance impressions which might be attributed differently for familiar identities. Therefore, we have reported these analyses in the Supplementary Materials (see Supplementary Tables 1 and 2).

Results and Discussion

Table 3 shows the mean fit of the data for familiar and unfamiliar faces as well as the fit for the two 'by chance' estimations (lower numbers reflect better fit).

The analysis shows a much closer fit for both types of faces compared to estimates of chance derived from two arbitrary recombinations of the same data. Furthermore, the familiar faces show a considerably closer fit than the unfamiliar faces, implying that images of the same familiar identity are located much closer to one another in this social

face space. A simplified example of the face space, making use of real rating data from the two face sets, is presented in Figure 3. Here, each point represents a different image and different colours represent different identities (left – unfamiliar identity, right – familiar identity). The figure illustrates that images of the familiar identity lie much closer together than the images of the unfamiliar identity. This suggests that social judgements are much less variable for familiar than unfamiliar faces. Once we are familiar with an identity and have formed a stable social impression, we are more likely to use this information as a cue when rating different images of the same identity.

Table 3

Mean Fit of Data as Well as Fit from the Chance Measures for Familiar and Unfamiliar Identities

	Familiar Faces		Unfamiliar Faces	
	With Scaling	No Scaling	With Scaling	No Scaling
Mean fit of data (SD)	.39 (.05)	.44 (.07)	.62 (.05)	.79 (.11)
Mean fit for Shuffle1 (SD)	.93 (.03)	1.48 (.12)	.93 (.03)	1.47 (.14)
Mean fit for Shuffle2 (SD)	.94 (03)	1.52 (.13)	.93 (.03)	1.50 (.14)

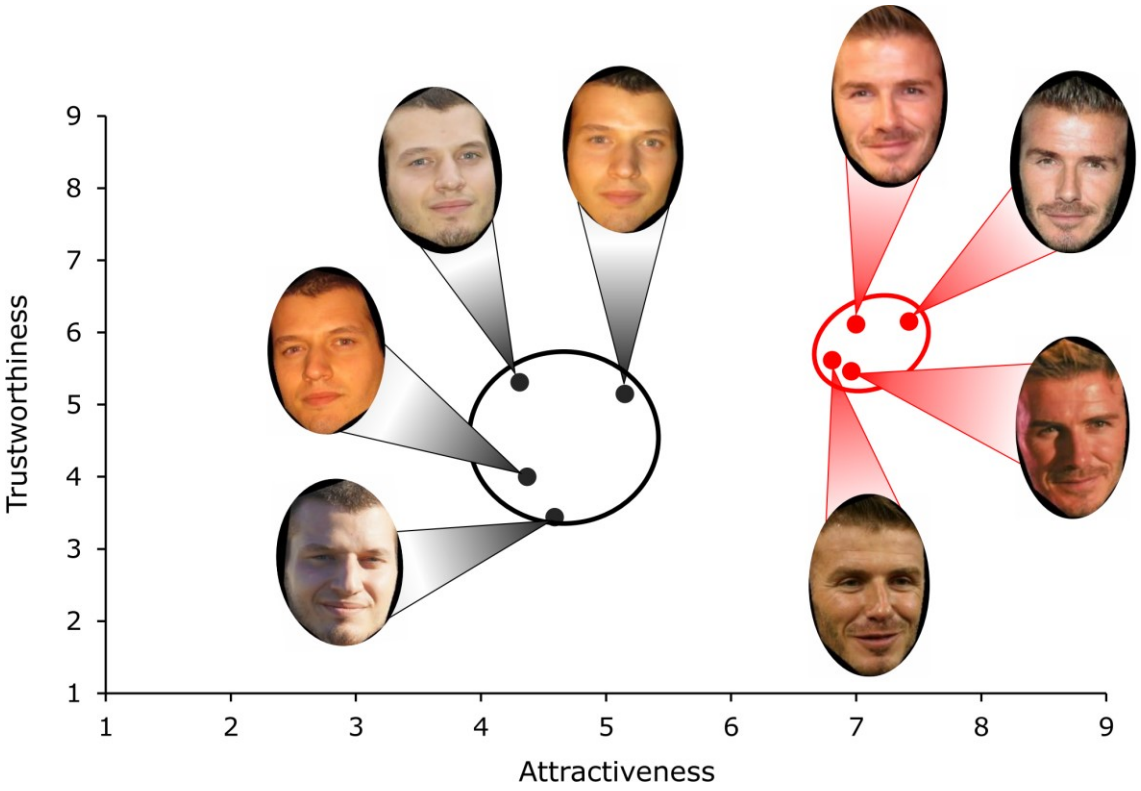

Figure 3. Example of the location of images in two-dimensional social attribute space for an unfamiliar (left) and familiar (right) identity. [Identity on the left has given permission for his images to be reproduced here. Images of the identity on the right (David Beckham) are labelled for reuse under creative commons licensing. See Acknowledgements for attributions].

Our Procrustes analysis shows a much closer fit for images of familiar identities. However, it does not eliminate the possibility that these images were simply less variable than those of the unfamiliar identities. Both the familiar and unfamiliar face sets contained images of celebrities which ensures that image quality and context are somewhat comparable. In order to address this issue further, we applied the same Procrustes analysis but this time based on the physical dimensions of the faces rather than their social evaluations.

Prior to the Procrustes fit analysis, we used principal components analysis (PCA) to extract the statistical properties of all 320 face images (160 familiar and 160 unfamiliar). First, we used the already aligned anatomical locations for each image (see Materials section, Study 1) to produce 82 *xy*-coordinates describing the shape of each face. The texture of the faces was then derived by calculating the average shape of the whole face set and then morphing the texture of each face to the average shape. This produced a texture vector of pixel intensities for each image. PCA was applied separately for face shape and texture to generate dimensions (referred to as eigenvectors or eigenfaces in the literature) which describe the way the faces vary. Each face can then be represented as a linear combination of these dimensions, which provides all images with a unique set of coefficients of mean zero, describing the position of each image within the face space created by the PCA. For the present analysis, we used the

first 50 shape and texture dimensions, therefore each image was coded as a set of 50 shape and 50 texture coefficients. For further details about the PCA procedure, see Burton, Kramer, Ritchie, and Jenkins (2016).

In order to estimate the similarity of familiar and unfamiliar images, we applied the same Procrustes analysis but this time based on the 50 face shape and texture dimensions. Therefore, if images of the same person are more physically similar to one another, they will lie closer together, whereas more different-looking images will be located further away from one another. The Procrustes analysis was applied separately for familiar and unfamiliar identities as well as for shape and texture components. Table 4 shows the mean fit of the data across familiarity and physical dimensions together with the two 'by chance' estimations. Again, lower fit values signify a better fit, i.e. high similarity between images of the same person.

The analysis demonstrates a very comparable fit between different images of familiar and unfamiliar identities. As with the Procrustes analysis on social attributes, we also applied the transformation without the scaling component (see Supplementary Table 3 for more details). Results showed the same pattern of results with no clear differences between familiar and unfamiliar identities. Such findings strengthen the interpretation of

our earlier findings as they demonstrate that the differences observed in Table 3 are not due to the physical differences between the familiar and the unfamiliar image sets.

Table 4

Mean Fit of Data as Well as Fit from the Chance Measures for Familiar and Unfamiliar Identities

	Familiar Faces		Unfamiliar Faces	
	Shape	Texture	Shape	Texture
Mean fit of data (SD)	.28 (01)	.32 (.02)	.31 (.01)	.36 (.02)
Mean fit for Shuffle1 (SD)	.88 (.01)	.88 (.01)	.88 (.01)	.88 (.01)
Mean fit for Shuffle2 (SD)	.34 (.01)	.37 (.02)	.37 (.02)	.39 (.02)

General Discussion

The present study aimed to examine the detailed structure of social attributions across gender and familiarity. We observed clear sex differences in the relationship between dominance and trustworthiness, with high dominance leading to more negative trustworthiness evaluations for female faces (Study 1). Further, we showed that social

ratings attributed to different images of the same familiar identity are much less variable than the ratings attributed to different images of the same unfamiliar identity (Study 2). Finally, we showed that even though the process of averaging leads to artefacts such as blurring and smoothing of skin texture, average images preserve some information about the overall social evaluation of those images (Study 1).

The observation of a strong negative relationship between trustworthiness and dominance for female faces only, fits well with social stereotype studies and Sutherland, Young, et al. (2015), who showed that counter-stereotypical (i.e. more dominant-looking) female faces were evaluated more negatively than stereotypical male and female faces, and even counter-stereotypical male faces. This implies that dominance might be interpreted differently for male and female identities and that people use different sets of cues when evaluating these traits. As most face evaluation models, however, are based on male faces only (Oosterhof & Todorov, 2008) or use male and female faces together (Walker & Vetter, 2009), further research is needed to support this assumption.

Another interesting finding is the strong positive relationship between distinctiveness and attractiveness for both male and female identities. Such results go against the vast literature on typicality, averageness and symmetry. In contrast to our findings, these

studies report that a more typical face is evaluated more favourably and perceived as more attractive. According to Thornhill and Gangestad (1993; 1999), a face close to the average signals the low probability of adverse genetic mutations being present, and further studies of cognitive processing have also established a link between averageness and ratings of attractiveness (Langlois et al., 1994). Instead, our results support Perrett (1994), who argues that while average (i.e. more typical and less distinctive) faces might be perceived as more attractive, there is more to attractiveness attribution than averageness. He found that exaggerating the shape of an attractive composite face made up of the 15 most attractive images in a face set lead to an increase in attractiveness ratings even though it changed the facial shape away from the average. Furthermore, such findings might reflect the variability in face shape and texture in the face database used for these studies. It is possible that certain identities were considered distinctive in the context of the present face set, however these same identities might not be as distinctive in the context of the general population.

In a first attempt to explore the relationship between physical and attributional variability for familiar identities, our Procrustes analysis demonstrated that different images of the same familiar identity were clustered much more closely together in a space of social dimensions. This implies that familiarity takes over differences in the physical properties of images in social evaluation, just as it does in face recognition,

whereas unfamiliar recognition is bound to the pictorial information in the face (Hancock et al., 2000). As demonstrated by Jenkins et al. (2011), this makes it difficult for us to cohere superficially different images of the same person into a single representation, or to 'tell people together' (Andrews et al., 2015). Jenkins et al. propose that as we become familiar with someone, we gain access to their idiosyncratic variability, reducing reliance on other irrelevant changes in different images of the same person. This has recently been demonstrated in face memory, where there is an advantage for familiar identities when testing memory for the person but an advantage for unfamiliar identities when testing memory for the specific image used to represent this person (Armann, Jenkins, & Burton, 2015). Our findings demonstrate that this familiarity mechanism extends to social evaluation. Here, the processing of all available physical information in images of unfamiliar identities leads to very different social ratings, whereas the focus on idiosyncratic information makes ratings attributed to images of familiar identities more consistent.

One caveat of the present studies is the use of celebrities' faces as stimuli. Collecting images of both internationally famous (used as familiar identities here) and locally famous (used as unfamiliar identities here) celebrities ensured that any differences found between these identities reflect familiarity rather than image quality, access to professional hair and makeup artists or image editing software. Nevertheless, it should

be noted that this can also restrict the amount of image variability compared to the faces we encounter in our everyday lives.

Finally, we show that a physical average image preserves some key social information about the person. This is despite evidence for averaging-related artefacts such as a blurring and smoothing of face texture, which could potentially influence social evaluation, especially with recent studies reporting a much stronger link between the physical properties of images and first impressions (Jenkins et al., 2011; Todorov & Porter, 2014). Our findings reveal another possible advantage of average images – they capture the most important idiosyncratic information about someone's face critical for both recognition and social evaluation. Addressing White et al.'s suggestion of implementing averages as an alternative to photo ID, it is important to know that socially-relevant cues are not lost through the process of image averaging. It should, nevertheless, be noted that the evidence for the accuracy of first impressions (i.e. whether they reflect any true personality characteristics of the person depicted) is limited at best (Todorov et al., 2015; although see Kramer & Ward, 2010). Therefore, we do not argue that the judgements attributed to average images present an advantage in terms of real world accuracy but rather that they preserve the most impression-relevant information in the face. Nevertheless, each participant rated all images of all

identities in the present studies, which could have artificially inflated the consistency between ratings attributed to different images of the same identity.

In summary, this paper shows clear gender differences in the relationships between the fundamental social evaluation dimensions. This contributes to our understanding of first impressions formation and implies that social traits might be interpreted differently when rating male and female faces. We also incorporate the effects of familiarity and demonstrate that it outweighs within-person variability in the context of social evaluation, just as it does in identity recognition (Jenkins et al., 2011). Such results support arguments for the differential processing of familiar and unfamiliar faces and challenge existing face recognition models.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.323262.

Attribution for Figure 1 (top row – left to right): Sgt. Bryson K. Jones [Public Domain], J. Chou [CC-BY-NC-ND 2.0], Michael Donovan [CC-BY-NC-SA 2.0], JCS [CC-BY-

SA 3.0], (bottom row – left to right): Eva Rinaldi [CC-BY-SA 2.0], Eva Rinaldi [CC-BY-SA 2.0], Eva Rinaldi [CC-BY-SA 2.0], Charlie Brewer [CC-BY-SA 2.0].

Attributions for Figure 3 (top centre – clockwise order): LCpl. Khoa Pelczar, United States Marine Corps [Public Domain]; AlexRoig2016, Àlex Roig Mangues [CC BY-SA 4.0]; The Democratic Alliance, David Beckham 2009 [CC BY-SA 2.0]; Paulblank, David-Beckham3 [CC BY 3.0].

References

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology, 68*, 10, 2041–2050. https://doi.org/10.1080/17470218.2014.1003949

Armann, R. G., Jenkins, R., & Burton, A. M. (2016). A familiarity disadvantage for remembering specific images of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 571–580. http://dx.doi.org/10.1037/xhp0000174

Baker, K. A., Laurence, S., & Mondloch, C. J. (2017). How does a newly encountered face become familiar? The effect of within-person variability on adults' and children's perception of identity. *Cognition, 161*, 19–30. https://doi.org/10.1016/j.cognition.2016.12.012

Balas, B., & Pearson, H. (2017). Intra-and extra-personal variability in person recognition. *Visual Cognition*, 1–14. http://dx.doi.org/10.1080/13506285.2016.1274809

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences, 104*(46), 17948–17953. doi: 10.1073/pnas.0705435104

Bruce, V. Henderson, Z. Greenwood, K. Hancock, P.J. B. Burton, A.M. Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*, 339–360. doi: 10.1037/1076-898X.5.4.339

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology, 66*, 8, 1467–1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Bruce, V., & Hancock, P. J. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science, 23*(1), 1–31. https://doi.org/10.1207/s15516709cog2301_1

Burton, A. M., Jenkins R., Hancock P. J. B., & White D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51,* 256–284. http://dx.doi.org/10.1016/j.cogpsych.2005.06.003

Burton, A. M., Kelly, S. W. W., & Bruce, V. (1998). Cross-domain repetition priming in person recognition. *Quarterly Journal of Experimental Psychology*, *51A*(3), 515–529. http://doi.org/10.1080/713755780

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40*, 202-223. doi:10.1111/cogs.12231

Burton, A. M., Wilson, S., Cowan, M, & Bruce, V. (1999). Face recognition in poor quality video: Evidence from security surveillance. *Psychological Science, 10,* 243–248. doi: 10.1111/1467-9280.00144

Dumas, R., & Teste, B. (2006). The influence of criminal facial stereotypes on juridical judgments. *Swiss Journal of Psychology, 65*(4), 237–244. http://dx.doi.org/10.1024/1421-0185.65.4.237

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

Flowe, H. D., & Humphries, J. E. (2011). An examination of criminal face bias in a random sample of police lineups. *Applied Cognitive Psychology, 25*(2), 265–273. doi: 10.1002/acp.1673

Gower, J. C. (1975). Generalized Procrustes analysis. *Psychometrika, 40*, 33–51. doi: 10.1007/BF02291478

Graham, J. R., Harvey, C. R., & Puri, M. (2016). A corporate beauty contest. *Management Science, Articles in Advance,* 1–13. http://dx.doi.org/10.1287/mnsc.2016.2484

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*, 330–337. https://doi.org/10.1016/S1364-6613(00)01519-9

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70.

Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R. (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology, 4*, 1–8. doi:10.3389/fpsyg.2013.00386

Jackson, D. A. (1995). PROTEST: A PROcrustean randomization TEST of community environment concordance. *Ecoscience, 2*, 297–303. http://dx.doi.org/10.1080/11956860.1995.11682297

Jenkins, R., & Burton, A. M. (2008). 100% accuracy in automatic face recognition. *Science, 319,* 435. doi: 10.1126/science.1149656

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society, B., 366,* 1671–1683. doi: 10.1098/rstb.2010.0379

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121,* 313–323. http://dx.doi.org/10.1016/j.cognition.2011.08.001

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory, 17*, 577–596. https://doi.org/10.1080/09658210902976969

Kramer, R. S. S., Jenkins, R., & Burton, A. M. (2017). InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior Research Methods, 49*, 2002–2011. https://doi.org/10.3758/s13428-016-0837-7

Kramer, R. S. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces. *PLoS ONE, 13*(8), e0202655. https://doi.org/10.1371/journal.pone.0202655

Kramer, R. S. S., Ritchie, K. L., & Burton, A. M. (2015). Viewers extract the mean from images of the same person: A route to face learning. *Journal of Vision*, *15*, 1–9. doi: 10.1167/15.4.1

Kramer, R. S. S., & Ward, R. (2010). Internal facial features are signals of personality and health. *The Quarterly Journal of Experimental Psychology, 63*, 2273–2287. https://doi.org/10.1080/17470211003770912

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, *1*, 115–121. doi: 10.1111/j.1467-9280.1990.tb00079.x

Langlois, J. H., Roggman, L. A., & Musselman, L. (1994). What is average and what is not average about attractive faces? *Psychological Science*, *4*, 215–221. doi: 10.1111/j.1467-9280.1994.tb00503.x

Laurence, S., & Mondloch, C. J. (2016). That's my teacher! Children's ability to recognize personally familiar and unfamiliar faces improves with age. *Journal of Experimental Child Psychology, 143*, 123–138. https://doi.org/10.1016/j.jecp.2015.09.030

Lippa, R. A. (2010). Gender differences in personality and interests: When, where, and why? *Social and Personality Psychology Compass, 4*, 1098–1110. doi: 10.1111/j.1751-9004.2010.00320.x

Little, A. C., & Hancock, P. J. (2002). The role of masculinity and distinctiveness in judgments of human male facial attractiveness. *British Journal of Psychology*, *93*, 451–464. doi: 10.1348/000712602761381349

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81–90. http://dx.doi.org/10.1037/0022-3514.52.1.81

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces aren't faces: Evidence from a matching task. *Memory & Cognition. 34,* 865–876. doi: 10.3758/BF03193433

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition, 128*, 56–63. http://dx.doi.org/10.1016/j.cognition.2013.03.006

Oldmeadow, J. A., Sutherland, C. A. M., & Young, A. W. (2013). Facial stereotype visualization through image averaging. *Social Psychological and Personality Science, 4*, 615–623. doi: 10.1177/1948550612469820

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

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

Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The Measurement of Meaning.* Urbana: University of Illinois Press.

Peres-Neto, P. R., & Jackson, D. A. (2001). How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia, 129*, 169–178. doi: 10.1007/s004420100720

Perrett, D. I. (1994). Facial shape and judgements. *Nature*, *368*, 239–242.

Ritchie, K. L., Mireku, M. O., & Kramer, R. S. S. (in press). Face averages and multiple images in a live matching task. *British Journal of Psychology*.

Ritchie, K. L., White, D., Kramer, R. S. S., Noyes, E., Jenkins, R., & Burton, A. M. (2018). Enhancing CCTV: Averages improve face identification from poor-quality images. *Applied Cognitive Psychology, 32*(6), 671–680. https://doi.org/10.1002/acp.3449

Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science, 19*(2), 109–111. doi: 10.1111/j.1467-9280.2008.02054.x

Rule, N. O., & Ambady, N. (2009). She's got the look: Inferences from female chief executive officers' faces predict their success. *Sex Roles, 61*(9), 644–652. doi: 10.1007/s11199-009-9658-9

Secord, P. F. (1959). Stereotyping and favorableness in the perception of Negro faces. *The Journal of Abnormal and Social Psychology*, *59*, 309–314. http://dx.doi.org/10.1037/h0042001

Sussman, A. B., Petkova, K., & Todorov, A. (2013). Competence ratings in US predict presidential election outcomes in Bulgaria. *Journal of Experimental Social Psychology, 49*(4), 771–775. http://dx.doi.org/10.1016/j.jesp.2013.02.003

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

Sutherland, C. A. M., Rowley, L. E., Amoaku, U. T., Daguzan, E., Kidd-Rossiter, K. A., Maceviciute, U., & Young, A. W. (2015). Personality judgments from everyday images of faces. *Frontiers in Psychology, 6*, 1616. http://doi.org/10.3389/fpsyg.2015.01616

Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (2015). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology*, *106*, 186–208. http://dx.doi.org/10.1016/j.cognition.2012.12.001

Schweinberger, S. R., Pfütze, E. M., & Sommer, W. (1995). Repetition priming and associative priming of face recognition: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(3), 722–736. http://dx.doi.org/10.1037/0278-7393.21.3.722

Thornhill, R., & Gangestad, S. W. (1993). Human facial beauty: Averageness, symmetry, and parasite resistance. *Human Nature*, *4*, 237–269. doi: 10.1007/BF02692201

Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in Cognitive Science*, *3*, 452–460. http://dx.doi.org/10.1016/S1364-6613(99)01403-5

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

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different images of the same person. *Psychological Science, 25*, 1404–1417. doi: 10.1177/0956797614532474

Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision, 9*, 1–13. doi: 10.1167/9.11.12

White, D., Burton, A. M., Jenkins, R., & Kemp, R. I. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied*, *20*, 166–173. http://dx.doi.org/10.1037/xap0000009

White, D., Sutherland, C. A., & Burton, A. L. (2017). Choosing face: The curse of self in profile image selection. *Cognitive Research: Principles and Implications, 2*, 23. https://doi.org/10.1186/s41235-017-0058-3

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

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*, 592–598. doi: 10.1111/j.1467-9280.2006.01750.x

Wilson, J., & Rule, N. (2016). Hypothetical sentencing decisions are associated with actual capital punishment outcomes: The role of trustworthiness. *Social Psychological and Personality Science*, *7,* 331–338. doi: 10.1177/1948550615624142

Young, A. W., Hellawell, D., & De Haan, E. H. (1988). Cross-domain semantic priming in normal subjects and a prosopagnosic patient. *The Quarterly Journal of*

Experimental Psychology Section A, 40(3), 561–580. https://doi.org/10.1080/02724988843000087

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

Supplementary Materials

Supplementary Table 1

Correlations Between Social Traits for Familiar Male and Female Identities.

	Male Identities				Female Identities			
	A	T	Dom	E	A	T	Dom	E
Attractiveness	–				–			
Trustworthiness	**.56****	–			**.76****	–		
Dominance	.53**	-.06	–		.32*	-.10	–	
Extraversion	.36**	.33*	.09	–	.15	.21	.01	–
Distinctiveness	**.43****	.40**	.45**	.37**	**.68****	.60**	.49**	.38**

N = 100, * $p < .05$, ** $p < .001$ (Holm-Bonferroni corrected). Bold, underlined correlations are reliably different between male and female identities.

Supplementary Table 2

Correlations between Ratings Attributed to the Physical Average and the Mean of the Four Exemplars for Familiar Faces.

	Male Faces		Female Faces	
	Mean of All Exemplars	Random Exemplar	Mean of All Exemplars	Random Exemplar
Attractiveness	.92	.87	.84	.89
Trustworthiness	.91	.82	.92	.86
Distinctiveness	.91	.76	.81	.85
Extraversion	.91	.78	.81	.69
Dominance	.85	.80	.91	.69

N = 20, all $p < .001$

Supplementary Table 3

Mean Fit of Data as Well as Fit from the Chance Measures for Familiar and Unfamiliar Identities with no Scaling

	Familiar Faces		Unfamiliar Faces	
	Shape	Texture	Shape	Texture
Mean fit of data (SD)	.30 (.03)	.36 (.03)	.34 (.02)	.40 (.04)
Mean fit for Shuffle1 (SD)	.89 (.01)	.88 (.01)	.88 (.01)	.88 (.01)
Mean fit for Shuffle2 (SD)	.37 (.02)	.43 (.04)	.42 (.02)	.48 (.04)