Within-Person Variability in Social Evaluation

Mila Mileva

Doctor of Philosophy

University of York

Psychology

May 2017

Abstract

When meeting someone for the first time, we not only extract a wealth of information about their age, gender, ethnicity, or mood, but we also evaluate them on social dimensions such as attractiveness, trustworthiness, or dominance. What makes these social inferences important and interesting is the fact that people agree with each other's evaluations and that they can influence our attitudes and behaviours, even if evidence for their accuracy is only limited. Existing face evaluation models focus on the identity level, arguing that a person is either, say, trustworthy or untrustworthy, regardless of the many different ways they might look. Recent evidence, however, suggests that images of the same person can vary just as much as images of different people, i.e. people rather have trustworthy- or untrustworthy-looking images of themselves. Here, I explore the spread and magnitude of such within-person variability in social evaluation. This is accomplished by sampling natural face variability and using images with different pose, emotional expression, lighting, etc. that are representative of real life social interactions. In addition to idiosyncratic variability, experiments described here aim to examine social evaluation across gender and familiarity as well as investigate the implications of trait inferences for face recognition. I then address social evaluation across modality, integrating visual information from the face and acoustic information from the voice. My findings show comparable within- and between-person variability in social ratings and demonstrate that idiosyncratic variability alone can bring about significant changes in trait attribution. This suggests that social evaluation depends on both identity and image properties. Finally, I demonstrate the automaticity of audiovisual integration in social evaluation and show that the relative contribution of face and voice cues is different for the two fundamental social dimensions. Ultimately, this brings us a step closer to understanding integrated person perception.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Thank you Mike, for all your support, understanding, and the opportunity to pursue something I feel so passionate about. I feel so incredibly fortunate to have been able to learn from you and work under your supervision. Thank you for your timely feedback and for always pushing me to do my best. Also, thank you for bringing me to Aberdeen, I had never been so alone, yet felt so at home.

I am also grateful to the ERC for funding my research.

Thank you to everyone in the lab, there was never a dull moment. I would have never done it without Kay's support and enthusiasm as well as Robin's relentless criticism and technical expertise. Also, thank you to Adam and Andrew for giving me a place to stay at the start of my PhD and for their good friendship. Last, but not least, thank you to Jenny for always making me feel good about myself and my work and Alice for helping me bring everything together.

Thank you to James and Dom for introducing me to the world of voices and making the last experimental chapter possible.

Most of all, thank you to my family for their unfailing love and support. Thank you to my brother for the endless supply of face images and to my dad for understanding that I love him even though I didn't call as much as I should have. To my mom, none of this would have been possible without your strength, determination, and selflessness.

Author's Declaration

I declare that this thesis is a presentation of original work and I am the sole author, with supervision from Professor Mike Burton. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. The research was funded by a studentship from the ERC.

Chapter 2: Procrustes analysis was performed with help from Dr Robin Kramer. Portions of this data were presented in a poster at the York Psychology Postgraduate Research Day, 2015.

Chapter 3: Principal components analysis was performed with help from Dr Robin Kramer. Experiments 3, 4, 5, and 6 appear as they have been submitted to the Journal of Experimental Social Psychology. Portions of this data were presented at the York Postgraduate Research Day, 2016, the Northeast Face and Person Perception Workshop, York, 2016 as well as at the British Psychological Society Cognitive Section Annual Conference, Barcelona, 2016.

Chapter 5: Experiments in this chapter were conducted in collaboration with James Tompkinson and Dr Dominic Watt from the Department of Language and Linguistic Science, University of York. Data from this chapter has been accepted in JEP: HPP. Portions of this data were presented at the internal seminars at the Department of Psychology, University of York on (2nd May 2017.

Mileva, M., Tompkinson, J. A., Watt, D., & Burton, A. M. (2017). Audiovisual Integration in Social Evaluation. *Journal of Experimental Psychology: Human Perception and Performance*. doi: 10.1037/xhp0000439

Chapter 1 – General Introduction

1.1. Introduction

The human face is an extremely rich stimulus and its accurate perceptual analysis is critical in the social world. It can provide us with a wealth of information about age, gender, race, emotional state, and identity (Bruce & Young, 1986). Relying solely on facial information, people readily form stable first impressions within a few milliseconds. Willis and Todorov (2006) presented participants with unfamiliar faces for 100, 500, or 1000 milliseconds and instructed them to rate each face for a number of social dimensions such as trustworthiness, attractiveness, and aggressiveness as well as to rate their confidence in these ratings. What they found was that 100 milliseconds were enough for participants to form these first impressions and any additional time had an effect on confidence ratings only. Further studies have explored this effect with more controlled methodologies and ever-shortening presentation time, demonstrating that a stable first impression can be formed in as little as 34 milliseconds (Bar, Neta, & Linz, 2006; Rule, Ambady, & Hallett, 2009; Todorov, Loehr, & Oosterhof, 2010; Todorov, Pakrashi, & Oosterhof, 2009). Furthermore, brain activity has also been shown to be able to track the attribution of social traits such as trustworthiness even when no such evaluation is required (Engell, Haxby, & Todorov, 2007) which goes on to imply that such processes are automatic and outside of conscious control.

Social attribution is characterised by a high level of agreement between observers and this consensus among people is something that was observed early on in psychology (Hollingsworth, 1922; Litterer, 1933). Since then a vast number of studies have replicated these findings (Albright et al., 1997; Zebrowitz & Montepare, 2008; Zebrowitz-McArthur & Berry, 1987), implying that there is some physical information in the face that observers use to inform their judgements. Further evidence comes from recent studies demonstrating such a consensus exists even in 3-to-4-year-olds and their ratings match these of adults (Cogsdill, Todorov, Spelke, & Banaji, 2014). The

high agreement in social judgements, however, does not necessarily mean that these attributions are accurate and reflect reality. Despite a large number of studies showing that participants can make accurate inferences from facial appearance about criminal behaviour (Porter, England, Juodis, ten Brinke, & Wilson, 2008; Rule, Krendl, Ivcevic, & Ambady, 2013) and political orientation (Rule & Ambady, 2010), Todorov, Olivola, Dotsch, and Mende-Siedlecki (2015) argue that these studies are not very well controlled and therefore fail to provide a true representation of social attribution accuracy. One such study, for example, reported that participants could correctly distinguish between a Republican and a Democrat political candidate 56% of the time, which was reliably above chance levels (Olivola, Sussman, Tsetsos, Kang, & Todorov, 2012). Once controlling for age, gender, and race, however, participants' accuracy dropped to chance levels (50.7%). Of greater relevance to the work described in this thesis is the assumption following from reports of high accuracy in social attribution – namely, that a person is either, say, trustworthy or untrustworthy, regardless of the different ways he or she may look. Recent studies, discussed in further detail below, demonstrate that there is a large variation in judgements of the same identity. Someone, whose photos are perceived as very highly trustworthy on average, may nevertheless have individual photos which are perceived as much less trustworthy (Todorov & Porter, 2014). While such findings question whether social attribution is an accurate reflection of reality, this does not necessarily mean that people do not act on their first impressions.

What follows in the sections below is a review of key and current face perception literature focused primarily on social evaluation. It starts by introducing the implications of social trait judgements, highlighting why the study of social evaluation is important. Then, I go on to discuss possible factors affecting face perception encompassing age, gender, emotional expressions, low-level image properties, distinctiveness, and familiarity. Next, the most influential face evaluation models and related techniques are introduced and evaluated on how they address, or rather fail to address, within-person variability and natural face variation. Finally, I present

research on the integration of facial and vocal cues and discuss social evaluation across modality.

1.2. Implications

Our first impressions can have an enormous role in our everyday life and have been shown to influence social outcomes in a variety of contexts, including dating preferences (Todorov et al., 2015), voting choices (Ballew & Todorov, 2007; Olivola, Funk, & Todorov, 2014; Olivola & Todorov, 2010a), eyewitness testimony (Mueller, Heesacker & Ross, 1984; Mueller, Thompson & Vogel, 1988), and sentencing decisions (Blair, Judd, & Chapleau, 2004; Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006). When it comes to mate choice, physical attractiveness is obviously a very important social attribute. However, other social traits have also been identified as critical in this context. In a study using data from a popular dating website, Olivola, et al. (2014, as cited in Todorov et al., 2015) explored the facial correlates of dating success. They found clear gender differences – while being perceived as fun and outgoing was positively correlated with dating success for men, being perceived as smart and serious was negatively correlated with dating success for women. What is important here is that these differences were still preserved even after controlling for physical attractiveness. Further insight into the importance of social judgements in mate choices also comes from Little, Burt, and Perrett (2006) who demonstrated that faces perceived to possess qualities, desired in a potential partner, were also rated as more attractive.

Ratings of competence from faces have been shown to be the most critical social attribute in the political context. A great number of studies have reported that political candidates who are perceived as more competent by unfamiliar raters are more likely to win elections and receive larger vote shares (Ballew & Todorov, 2007; Sussman, Petkova, & Todorov, 2013; see Olivola & Todorov, 2010a for a review). It is interesting to note, however, that some cultural differences have been observed in this context as competence judgements have been shown to be better predictors of American rather than Korean elections (Na, Kim, Oh, Choi, & O'Toole, 2015). In the business

environment it has been reported that CEOs who are perceived as more competent and dominant receive larger salaries (Rule & Ambady, 2008, 2009), even though they sadly do not perform any different from other less competent-looking CEOs (Graham, Harvey, & Puri, 2016). Moreover, comparing the influence of social attributes on salaries of senior managers and lower shop-floor managers, Fruhen, Watkins, and Jones (2015) have shown that attractiveness is related to managerial pay awards in the lower shop-floor level, while perceived trustworthiness and dominance were correlated with senior managerial pay awards.

The most worrying influences of social attribute judgements are in relation to eyewitness testimony, sentencing decisions, and punishment severity. For example, it has been shown that people perceived to possess stereotypically criminal-looking faces are more likely to be picked out from a police line-up and consequently face trial (Flowe & Humphries, 2011). Moreover, Afro-centric facial features have also been linked to harsher sentences (Blair et al., 2004) and a higher likelihood to receive death penalty (Eberhardt et al., 2006). Further still, Dumas and Teste (2006) found that defendants with faces that fit the stereotype of the committed crime, they are being tried for, were more likely to be pronounced as guilty even with less substantial evidence. In a very recent and alarming study, Wilson and Rule (2016) asked participants to rate images of death row inmates for trustworthiness and found that perceivers rated the inmates sentenced to death as less trustworthy than inmates sentenced to life, replicating their previous findings (Wilson & Rule, 2015). Critically, perceived face trustworthiness (and not Afro-centricity, attractiveness or babyfaceness, which are other traits that have been found to affect sentencing) was the only trait that accounted for this relationship. Participants were then presented with face images of already convicted criminals and asked to assign a *life without parole* or a *death* sentence without any further information. These hypothetical sentencing decisions matched the actual sentences received in court, e.g. people who were sentenced to death, were also more likely to receive a hypothetical death sentence based on their face image only.

1.3. Factors Affecting Social Attribution

Social evaluation is affected by a range of perceptual and higher-order factors, including image contrast, face redness, and distinctiveness (Russell, 2003; Sofer, Dotsch, Wigboldus, & Todorov, 2015; Stephen & Perrett, 2015). Moreover, it has been associated with a number of overgeneralisation processes where cues related to age, gender, and emotion are used to infer stable personality attributes. Overgeneralisation can be traced back to work by Secord (1959) on social categorisation, arguing that first impressions results from firstly assigning a category to a face and then using category-associated information to evaluate it. The most systematic contemporary research on these overgeneralisation effects has been conducted by Zebrowitz and colleagues (see Zebrowitz, 2011 for a review).

Age and gender overgeneralisation

Starting with age, Zebrowitz has shown that babyfaced adults are attributed childlike characteristics. The morphological characteristics of such faces include large eyes, lighter skin and hair, rounder face, and lower vertical placement of features, which results in higher brows and forehead (Zebrowitz & Montepare, 2008). Faces possessing these features are universally perceived as more submissive, warm, honest, and naïve as well as physically and intellectually inferior (Montepare & Zebrowitz, 1998; Zebrowitz, 1997). They are also seen as more helping, caring and in need of protection (Berry & McArthur, 1986). Faces with more mature features, on the other hand, are associated with lower levels of attractiveness, health, and warmth and perceived as more likely to be experts and command respect (Montepare & Zebrowitz, 2002). Not only does age generalisation guide social evaluation, but it also leads to related social outcomes. Regardless of their actual age and sex, people with babyish features are more likely to be exonerated when charged with intentional crimes, but more likely to be charged with negligence than people with more mature features (Montepare & Zebrowitz, 1998). Montepare and Zebrowitz draw on ethological evidence (Eibl-Eibesfeldt, 1989) and argue that this process is triggered by a strong intrinsic prepared response to babyish facial cues guided by the evolutionary

importance to respond to such cues. This hypothesis is further supported by the high levels of agreement in social evaluation across cultures (Zebrowitz, Montepare & Lee, 1993) and studies showing a similar overgeneralisation effect even in infants and young children (Kramer, Zebrowitz, San Giovanni, & Sherak, 1995; Montepare & Zebrowitz-McArthur, 1989).

Gender overgeneralisation is a process closely related to perceived age from faces (Zebrowitz, 1997). As female faces are more likely to preserve youthful facial characteristics as they age, they are evaluated similarly to babyish faces, whereas male faces are generally ascribed attributes related to mature-faced individuals (Enlow & Hans, 1996). Consequently, female faces are perceived as more submissive and caring, while male faces are seen as more dominant, capable, and intelligent. Such effects further apply to faces with gender counter-stereotypical characteristics where less feminine female faces are evaluated as less attractive, intelligent, and sociable (Cunningham, 1986) and less masculine male faces are perceived as less dominant and healthy (Luevano & Zebrowitz, 2007). Apart from overgeneralisation, gender is a particularly salient cue for dominance attribution where male faces are generally perceived as more dominant than female faces (Boothroyd, Jones, Burt, & Perrett, 2007; Buckingham et al., 2006, see Figure 1.1). The distinction between masculinity and femininity is assigned a different role in first impression models, which separate social evaluation into two fundamental dimensions – valence and dominance. While Oosterhof and Todorov (2008) argue that masculinity/femininity is mostly related to perceptions of dominance in face evaluation, other general social evaluation models relate valence to femininity and dominance to masculinity (Cuddy, Fiske, & Glick, 2008; Prentice & Carranza, 2002; Wiggins, 1979). As most first impression studies have been focused on identifying the facial information responsible for social attribution, little is known about the relationship between social traits across gender. Sutherland, Young, Mootz, & Oldmeadow (2015), for example, explored ratings of male and female faces with gender stereotypical and counter-stereotypical characteristics. She showed a negative relationship between dominance and both attractiveness

and trustworthiness for female faces but no such pattern for male faces, demonstrating clear gender-based differences in social evaluation.

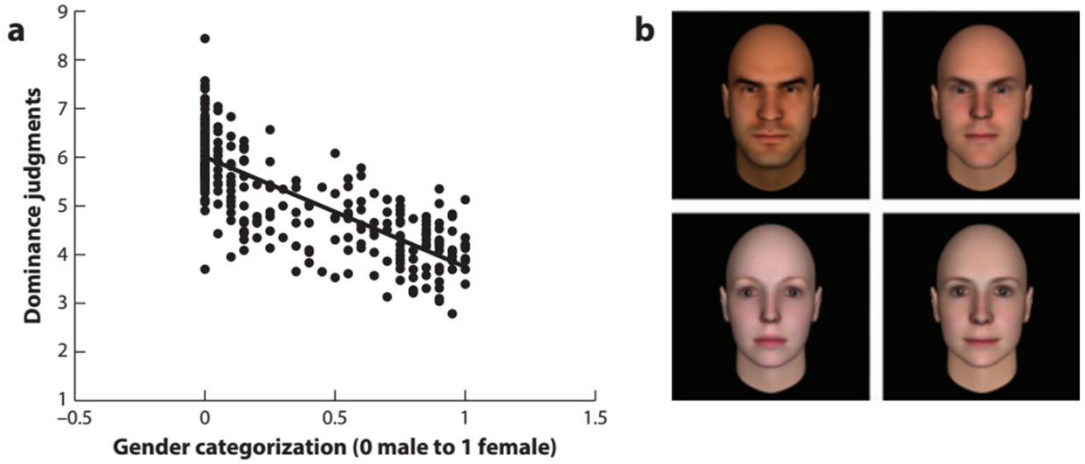

Figure 1.1. Relationship between gender and dominance attribution from Todorov et al. (2015). (a) presents a scatterplot of dominance ratings and gender categorisation. In (b), faces presented in the top row are the two faces rated as the most dominant and faces in the bottom row are the ones rated as the least dominant.

Attractiveness overgeneralisation

Similarly to younger-looking and feminine faces, people with attractive faces are also evaluated more positively (Eagly, Ashmore, Makhijani, & Longo, 1991; Langlois et al., 2000). Out of all social attributes, attractiveness has received the most research attention in first impressions. It has been consistently associated with a range of positive dimensions in what is described as the 'halo effect' (Dion, Berscheid, & Walster, 1972). Attractive faces are perceived as more intelligent, outgoing, capable, socially competent, and healthier (Feingold, 1992; Zebrowitz, Hall, Murphy, & Rhodes, 2002), whereas those perceived as less attractive are also seen as more dishonest, antisocial, psychologically unstable, and less willing to cooperate (Mulford, Orbell, Shatto, & Stockard, 1998). In addition to perception, this 'halo effect' seems to apply to behaviour as studies have demonstrated preferential treatment of attractive people in a range of contexts such as court decisions,

occupational settings, and interpersonal relations (Langlois et al., 2000; Little, Burt, & Perrett, 2006; Zebrowitz, 1997). Moreover, evidence for the existence of this effect across cultures and ages suggests that it reflects a global, rather than an arbitrary mechanism (Cunningham, Roberts, & Barbee, 1995; Ramsay, Langlois, Hoss, Rubenstein, & Griffin, 2004). Zebrowitz and Rhodes (2004) explain attractiveness overgeneralisation with the anomalous face hypothesis which argues that the importance of identifying unhealthy people with 'bad genes' has attuned us to detect facial markers of low fitness. As unattractive faces are more likely to possess such characteristics, they trigger the same negative attitude as faces of unfit or unhealthy individuals. Following this argument, some even suggest that the 'halo effect' is not driven by the perception that 'beautiful is good', as was originally thought, but by the perception that 'ugly is bad' (Griffin & Langlois, 2006). Support for this hypothesis comes from a connectionist neural activation model, trained to respond to anomalous faces. It demonstrated that facial metrics of unattractive faces corresponded more closely to those of anomalous faces and that faces structurally similar to anomalous ones were associated with more negative social impressions (Zebrowitz, Fellous, Mignault, & Andreoletti, 2003).

Emotion overgeneralisation

Emotional expressions and emotion overgeneralisation possibly describe the most influential and extensively studied factor in social evaluation. In terms of emotional expressions, faces displaying positive emotions, such as happiness, are evaluated more favourably and considered more trustworthy, friendly, kind, and easy-going (Krumhuber et al., 2007). Faces that possess features of anger (e.g. low brows, thin lips and withdrawn corners of the mouth), however, are seen as more dominant, threatening, and aggressive (Hess, Blairy, & Kleck, 2000; Montepare & Dobish, 2003). This describes the emotion overgeneralisation hypothesis where emotional information in the face is not only interpreted as evidence of momentary affective state but also of stable personality characteristics (see Zebrowitz, Kikuchi, & Fellous, 2010 for a review). What is most striking about the influence of emotion on social evaluation is that it is generalised to neutral

faces, which subtly resemble a specific emotional expression (Said, Sebe, & Todorov, 2009). Some argue that this process is driven by the similarity of face expression related to both temporary emotional states and enduring traits (e.g. aggressiveness and anger; Zebrowitz & Montepare, 2008). This hypothesis is supported by studies showing that neutral faces, rated as looking happier by one sample of participants are rated more favourably by another sample of participants. Similarly, neutral faces rated as looking angrier are attributed more negative social traits and are also perceived as more dominant (Montepare & Dobish, 2003). A competing explanation, however, suggests that these similarities between traits and emotions are based on semantic, rather than purely visual information (Schneider, 1973).

In an attempt to distinguish between these two competing mechanisms Said, Sebe, and Todorov (2009) used a Bayesian network classifier, trained specifically on emotional expressions, to detect even the most subtle emotional resemblance in neutral faces, thus eliminating any semantic information. The network compared the position of pre-specified face landmarks in emotional and neutral faces and reported the probability that the neutral face resembled the six basic emotions (happiness, surprise, anger, disgust, sadness and fear). These probability values were then related to social attribute ratings provided by human participants with the idea that classification probabilities will predict trait judgements if emotion overgeneralisation is based on purely visual similarities between emotions and personality traits. Results showed a significant positive correlation between positive social traits (caring, responsible, and sociable) and the probability of classifying faces as happy, and a significant negative correlation with negative traits such as aggressiveness. The probability of classifying faces as angry, on the other hand, was positively correlated with perceptions of dominance, unhappiness, and meanness and a resemblance to fear was interpreted as being submissive, insecure, irresponsible, and less intelligent (see Figure 1.2 for details on all social attributes). Such findings support the structural similarity argument and suggest that this process is due to an overgeneralisation of emotion recognition systems in the brain.

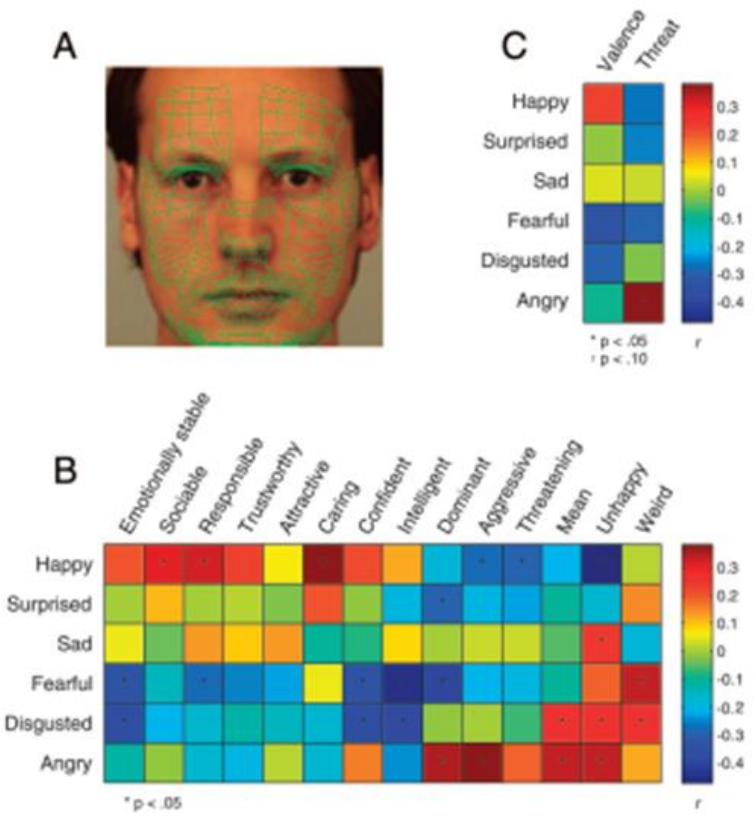

Figure 1.2. Relationship between social judgements and classified emotion probabilities as reported in Said et al., 2009. (A) is an example of the facial landmark used to detect subtle changes in emotion. (B) presents the correlation between classifier probabilities and each specific trait and (C) shows the correlation between these probabilities and the two fundamental social evaluation dimensions (valence & threat).

Physical image properties

In addition to overgeneralisation, there is also a range of image properties that have been shown to influence the way images of different people are perceived (Carre & McCormick, 2008; Menzel, Hayn-Leichsenring, Langner, Wiese, & Redies, 2015; Perrett, 2010). These measures encompass colour and texture differences such as contrast, brightness, and redness in the face as well as face shape differences such as facial width-to-height ratio (FWHR).

Colour changes in the face occur naturally as a result of both stable conditions (physical health) and transient states (emotions) and while these changes might be subtle, research has shown that people seem to be attuned to detect them (Changizi, Zhang, & Shimojo, 2006; Tan & Stephen, 2013). A variation in face colouring that has been of particular interest to face evaluation research is the amount of redness in the facial skin (see Stephen & Perrett, 2015 for a review). It has been suggested that redness in the face can be a result of blood oxygenation in the skin which is an indicator of good physical health (Stephen, Coetzee, Law Smith, & Perrett, 2009) as well as of higher intake of fruit and vegetables, due to carotenoid colouring (Stephen, Coetzee, & Perrett, 2011). A study by Stephen, Law Smith, Stirrat, and Perrett (2009) demonstrates our sensitivity to these colour differences by presenting male participants with images of female faces and asking them to manipulate the colour of the faces in digital photography in order to make them look as healthy as possible. Results of the study showed that men consistently increased the amount of redness in the face to make women appear healthier. As perceptions of good health are cues to attractiveness (Weeden & Sabini, 2005), it is possible that redness in the facial skin is a predictor of attractiveness attribution and there is some evidence to support their association (Pazda, Thorstenson, Elliot, & Perrett, 2016; Re, Whitehead, Xiao, & Perrett, 2011; see Figure 1.3 for stimuli examples). Furthermore, facial redness seems to influence other social evaluation dimensions with research demonstrating that increased redness in the face also enhances the perception of dominance (Stephen, Oldham, Perrett, & Barton, 2012).

The effect of image contrast on the perception of attractiveness is brought about by the use of facial cosmetics. Make-up application has been shown to improve the homogeneity of facial skin tone as well as increase the contrast between the features of the face (Jones, Russell, & Ward, 2015). This increase in contrast has been consistently associated with higher ratings of attractiveness (Porcheron, Mauger, & Russell, 2013) and dominance (only for female participants rating female faces, Mileva, Jones, Russell, & Little, 2016). Apart from the use of cosmetics, contrast seems to have a different effect on attractiveness evaluation across the two genders, with increased

contrast leading to higher ratings of attractiveness for female faces and lower contrast leading to higher ratings of attractiveness for male faces (Russell, 2003). The level of brightness in an image has also been associated with differences in social evaluation (Valdez & Mehrabian, 1994). Lakens, Fockenberg, Lemmens, Ham, and Midden (2013), for example, showed that neutral images were evaluated more positively when their lightness was increased and this association was also reported with emotional stimuli where smiling faces were perceived as brighter in colour than faces with a frown, even though image brightness was held constant for all images (Song, Vonasch, Meier, & Bargh, 2012). As some of those studies investigate general, rather than face-specific social evaluation, however, it is not clear whether these image properties are valid cues to social face perception.

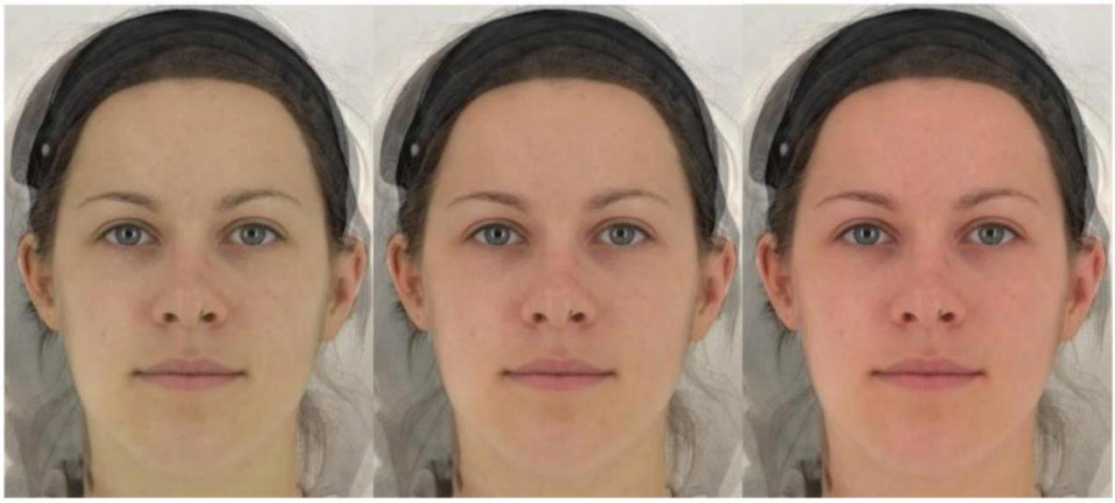

Figure 1.3. Example of the face stimuli used by Pazda et al. (2016). Faces were manipulated on the CIELAB a* (redness) colour axis by -5 units (left) and +5 units (right).

Another low-level image property that could have an influence on social face evaluation is spatial frequency power, as measured by the Fourier slope of the image. This measure is widely used in natural scene perception where a variety of aesthetically pleasing images have been shown to possess specific patterns in their spatial frequency distribution (Graham & Field, 2007; Redies, Hasenstein, & Denzler, 2007). The Fourier slope of an image measures the amount of fine detail (high spatial frequencies) and coarse

structure (low spatial frequencies) that it contains with shallower slopes indicating a larger proportion of high spatial frequencies and steeper slopes indicating a larger proportion of low spatial frequencies. Analyses of substantial datasets of aesthetically pleasing natural scenes and artworks demonstrate that they possess a slope of approximately -2, whereas face images present with relatively steeper slopes (of approximately -3.5, Koch, Denzler, & Redies, 2010). This suggests that faces with a shallower slope (closer to -2) might be perceived as more aesthetically pleasing and attractive. In a series of experiments, Menzel et al. (2015) manipulated the slope of face images as well as the slope of the image background and found that faces with shallower slope (closer to -2) and faces presented on a background with a slope closer to -2 were consistently rated as more attractive than faces with steeper slopes, supporting the role of spatial frequency distribution in social face evaluation (see Figure 1.4 for a manipulation example).

Figure 1.4. Faces with manipulated Fourier slope used in Menzel et al. (2015).

A face metric related to facial shape, rather than texture, that has been associated with first impressions is facial width-to-height ratio (FWHR, defined as the bizygomatic width divided by the upper facial height). Variations in this physical measure have been shown to predict both ratings of dominance and reactive aggression as well as actual aggressive behaviour (Carre & McCormick, 2008; Carre, McCormick, & Mondloch, 2009). Furthermore, Stirrat and Perrett (2010) have also reported that FWHR predicted 16% of the variance in trustworthiness decisions with wider faces

trusted less. In a subsequent experiment, the FWHR of faces was artificially increased or decreased and participants were significantly more likely to select images with lower ratio as more trustworthy. It should be noted, however, that not all studies report such a relationship between FWHR and aggression (Gómez-Valdés et al., 2013; Özener, 2011). Large within-person differences have also been reported for this face metric, which is especially pertinent to the experiments in this thesis, as they are focused on within-person variability (Kramer, 2016). Figure 1.5 shows how FWHR can change within a single identity.

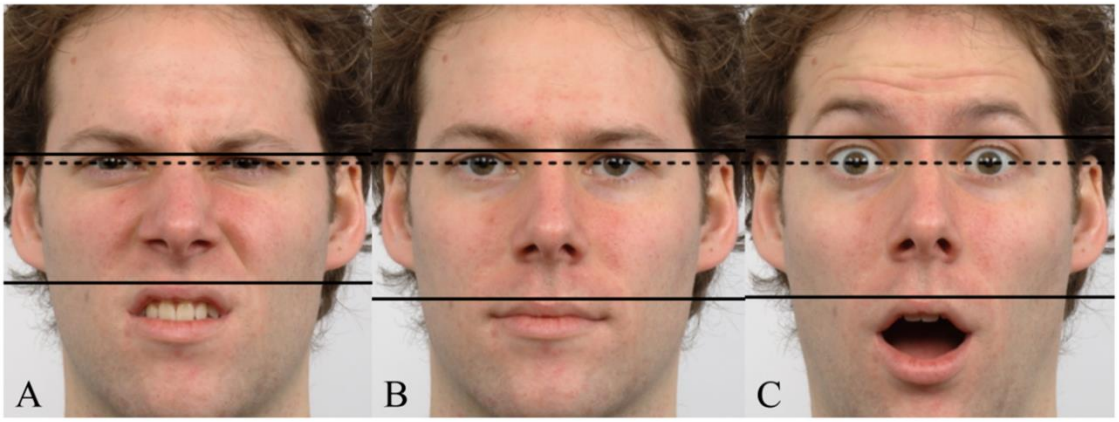

Figure 1.5. Facial width to height ratio within a single identity across different emotional expressions (Kramer, 2016).

Higher-level factors

Familiarity

In addition to overgeneralisation and low-level physical properties, social evaluation is also shaped by higher-order factors, such as familiarity and distinctiveness (Sofer et al., 2015; Zebrowitz & Montepare, 2008). Familiarity is one of the most important factors in face recognition, with the consistent finding that familiar faces are recognised quicker and more accurately than unfamiliar faces (Burton, White, & McNeill, 2010; Clutterbuck & Johnston, 2002; 2005). This has been demonstrated in face recognition tasks, which test face memory (Bruce, 1982; Ellis, 1981), as well as in purely perceptual face matching tasks (Bruce, 1986; Bruce, Henderson,

Newman, & Burton, 2001). Studies investigating this familiarity advantage suggest familiar and unfamiliar faces are processed in a qualitatively different way. Accurate recognition of familiar people, for example, has been shown to depend mostly on the internal features of the face (i.e. the parts inside the face outline; Ellis, Shepherd, & Davies, 1979; Young, Hay, McWeeny, Flude, & Ellis, 1985), whereas external features such as hair and face shape are more important for unfamiliar recognition (Bonner, Burton, & Bruce, 2003; Bruce et al., 1999).

Despite the fact that familiarity seems irrelevant in the context of zero-acquaintance impressions, it is still possible that the distinct processing of familiar and unfamiliar faces affects social evaluation. Building upon the association between unfamiliar faces and external features, for example, there is evidence for the significance of both face shape and hair in social evaluation. Face shape can be related to the width-to-height ratio, which has been shown to affect the perception of dominance and masculinity (Carre & McCormick, 2008; Carre, McCormick, & Mondloch, 2009) as well as to age overgeneralisation, where a rounder face is associated with babyfaceness and therefore evaluated more positively than more angular faces (Montepare & Zebrowitz, 1998; Zebrowitz, 1997). As the most influential face evaluation model uses computer-generated faces with no hair (Oosterhof & Todorov, 2008), the importance of hair for first impressions has not been extensively studied. Nevertheless, there is some evidence that grooming and hairstyle might convey social rank and self-esteem (Kaiser, 1985) and that lighter hair colour is also a marker of babyish faces (Cunningham, Barbee, & Pike, 1990). Cunningham further reported an association between hair length and attractiveness ratings of male faces and also linked baldness to evaluation of mature-looking individuals (Muscarella & Cunningham, 1996).

In what has been referred to as the mere exposure effect (Zajonc, 1968), a sense of familiarity leads to more positive and favourable social evaluation. This is supported by studies reporting a positive relationship between familiarity and both attractiveness and trustworthiness ratings (Peskin & Newell, 2004; Zebrowitz, Bronstad, & Lee, 2007) as well as by face

recognition studies where previously seen faces are rated as more attractive than 'new' faces (Rhodes, Halberstadt, & Brajkovich, 2001; Rhodes, Halberstadt, Jeffery, & Palermo, 2005). Zebrowitz accounts for this effect with the familiar face overgeneralisation hypothesis, which highlights the evolutionary value of being able to differentiate between known identities and strangers (Zebrowitz & Collins, 1997; Zebrowitz & Montepare, 2008). Thus, the more someone resembles a known identity, the more positively he/she will be evaluated. Consistent with this explanation, Kraus and Chen (2010) reported more positive ratings attributed to faces resembling one's significant others. This was further extended by Verosky and Todorov (2010a) who manipulated faces to subtly resemble identities previously associated with either negative or positive behaviours and showed more a positive evaluation of the latter faces, even when participants were specifically instructed to ignore any resemblance effects (Verosky & Todorov, 2013). Moreover, the same is true for faces manipulated to resemble participants' own faces (Verosky & Todorov, 2010b). DeBruine (2005), for example, demonstrated that such faces are seen as more trustworthy and Bailenson, Iyengar, Yee, and Collins (2009) showed a preference for faces of political candidates that have been altered to resemble participants' own faces.

Face typicality

As early as the 1880s Sir Francis Galton (1883) noted a link between familiar face overgeneralisation and face typicality with the potential to affect social evaluation. According to him averaging faces with the same nationality will result in a national "ideal" (typical) face, which might be the most consensually familiar face in a population and therefore serve as a comparison standard when evaluating novel faces on socially-important dimensions. Faces located away from this "ideal" face in face space, therefore, would be evaluated more negatively than faces closer to the typical nation face. Contemporary face evaluation studies provide support for Galton's argument with studies reporting a positive relationship between face typicality and attractiveness (DeBruine, Jones, Unger, Little, & Feinberg, 2007; Langlois, Roggman, & Musselman, 1994; Said & Todorov, 2011). Langlois and Roggman (1990), for example, showed higher attractiveness

ratings for an average face made up of 32 faces, compared to averages made up of subsets of those images or most individual faces. This was further confirmed in a meta-analysis reporting a medium-to-large effect of typicality on attractiveness attribution (Rhodes, 2006). Perrett (1994), however, reported some inconsistent results where 60-image averages were rated as less attractive than an average of the 15 most attractive faces in the same set.

As face evaluation models argue that trustworthiness, rather than attractiveness, represents one of the fundamental dimensions of social judgements from faces (Oosterhof & Todorov, 2008), Sofer et al. (2015) explored the effect of distance from the typical face on the perception of both attractiveness and trustworthiness. They used a typical face and a composite face made up from images with high attractiveness ratings to create an 11-image continuum with the typical face as a mid-point and by either adding or subtracting the difference in shape and texture between the typical and the attractive composite face (see Figure 6 for an example). These images were then rated for attractiveness and trustworthiness revealing clear differences between the typicality-attractiveness and typicality-trustworthiness relationships (see Figure 1.6). While the typical face was perceived as the most trustworthy, attractiveness ratings kept increasing past the typical face and towards the attractive composite. Consistent with Perrett (1994), this argues against the linear relationship between attractiveness and typicality and also demonstrates the importance of typicality for one of the fundamental social evaluation dimensions – trustworthiness.

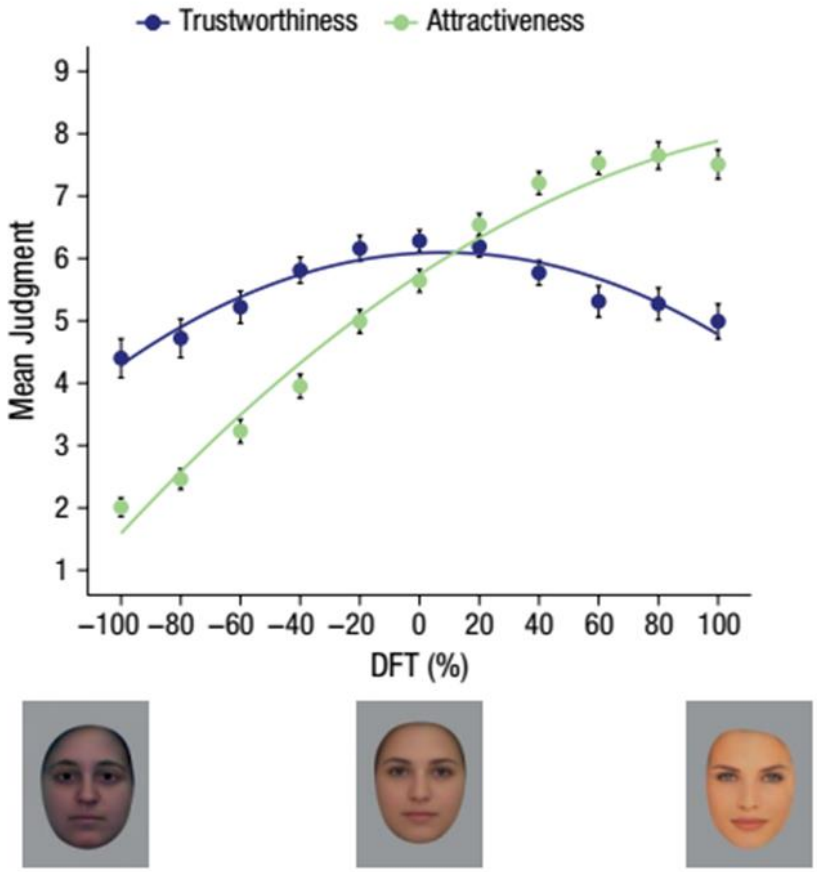

Figure 1.6. Data from Sofer et al. (2015) on the relationship between trait judgements and distance from the typical face (DFT).

1.4. Principal Components Analysis

Principal components analysis (PCA) is one of the most common statistical techniques in face perception and recognition. It has been used to model face similarity effects, face distinctiveness, the other-race effect, and emotional expressions (Calder, Burton, Miller, Young, & Akamatsu, 2001; Hancock, Bruce, & Burton, 1998; Hancock, Burton, & Bruce, 1996; O'Toole, Deffenbacher, Valentin, & Abdi, 1994). What makes this technique appropriate and useful in face perception is its ability to represent the variability of multidimensional data in few dimensions (also referred to as "eigenfaces" in the literature). This was demonstrated in the original work of Kirby and Sirovich (1990) as well as Turk and Pentland (1991) who showed

that face images can be reconstructed using as few as 50 eigenfaces compared to many thousands required in a pixel-by-pixel representation (Burton, Bruce, & Hancock, 1999). Another advantage of this technique is that it encodes faces in a holistic manner which echoes many proposals for human face perception (Young, Hellawell, & Hay, 1987).

Eigenfaces are generated once a set of face images is subjected to PCA. The images are then re-coded or reconstructed in the space of a subset of these eigenfaces, assigning each image a unique set of coefficients which act as its signature. A common practice in many PCA approaches is to employ some form of shape-normalisation which separates face shape and texture. In this context, face texture codes for information on the face surface, colour, reflectance and lighting. This process requires placing a standard grid on each face and altering it by hand to align with key landmark points (e.g., positions of inner and outer corners of the eyes or corners of the mouth, see Figure 1.7A for an example grid). The shape and texture of the image are then separated by morphing the image to a standard shape which usually is the average shape of all faces in the set. The images produced following this procedure, called “shape-free faces” (Craw, 1995), are then used for texture PCA (see Figure 1.7C for examples). The shape component, therefore, codes the original position of the points in the grid while the texture component codes the pixel intensities in its standardised shape. Following this separation PCA is applied independently to the shapes and textures of the images and each image is then assigned a unique set of shape and texture coefficients which are used to represent them in this low-dimensional space usually using the early eigenfaces of shape and texture.

In face evaluation, PCA is usually applied to databases containing only one image of each identity which aims to identify the dimensions along which faces of different people vary. Experiments in this thesis, however, aim to extract idiosyncratic variability, or the underlying physical dimensions along which images of the same identity may vary. This is accomplished by sampling many images of that same person spanning long- and short-term changes in the face (e.g., age or emotional expression) as well as world

variability due to camera angle or lighting. A key requirement to achieve this is the use of naturally occurring (or "ambient") images which are not taken under controlled conditions or varied in a systematic way. Adopting this approach makes it possible to span the space of each person's variability and provide a much deeper understanding of the entire visual range of that particular person's face.

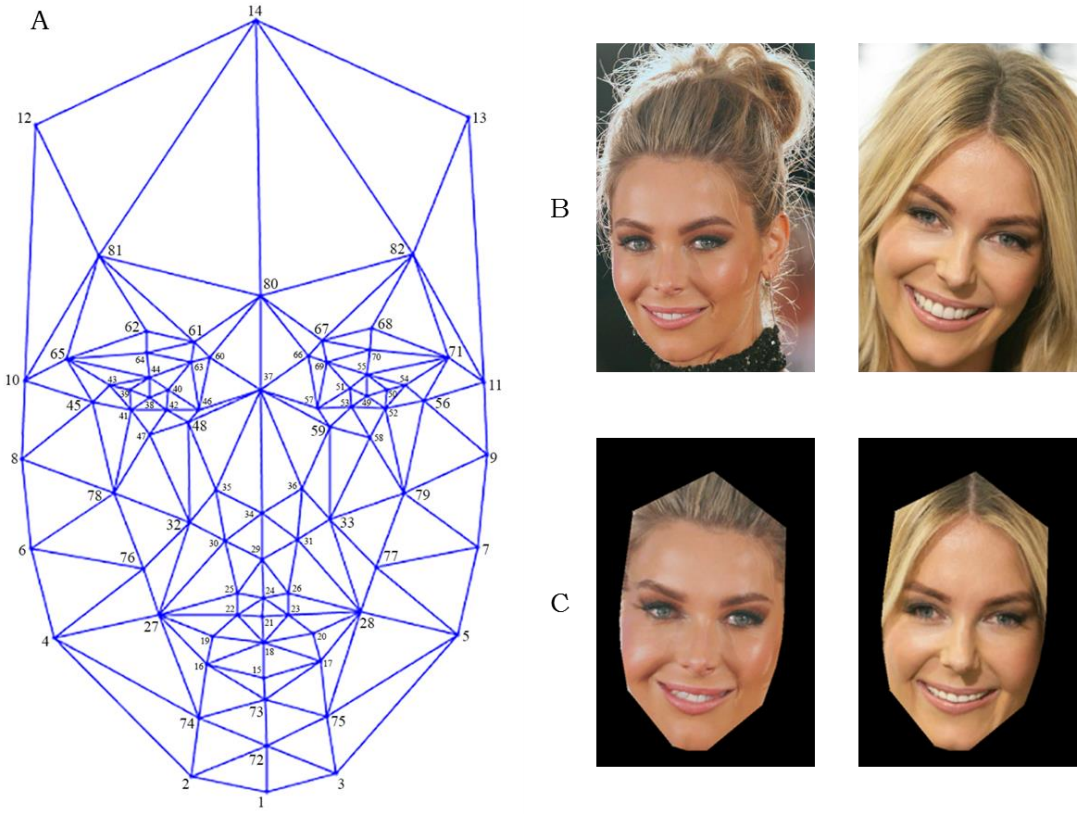

Figure 1.7. (A) shows an example of a grid with landmark positions used to extract face shape (Kramer, Jenkins, & Burton, 2016). (B) and (C) show original images and their shape-free textures respectively.

Utilising this within-person approach Burton, Jenkins, and Schweinberger (2011) performed PCA on 48 ambient images of individual people and examined the statistical properties of their personal shape and texture dimensions. Their analyses revealed that the early components covered superficial variability such as direction of light or camera angle (see Figure 1.8). As these early components explain the most variance in the image set,

this implies that the biggest differences between images of the same person are caused by changes in the world rather than changes in the face. Such results fit well with the existing between-person PCA literature as similar dimensions are commonly extracted from multiple identity sets. Within-person PCA, however, allows us to go a step further from world and general face variability and Burton et al. (2011) demonstrate that this idiosyncratic variability starts to emerge as early as the third component (which in the case of Harrison Ford, below, codes for changes in emotional expressions). This demonstrates that PCA can be used to extract information not only about the ways faces of different people vary but also information about identity-specific variability.

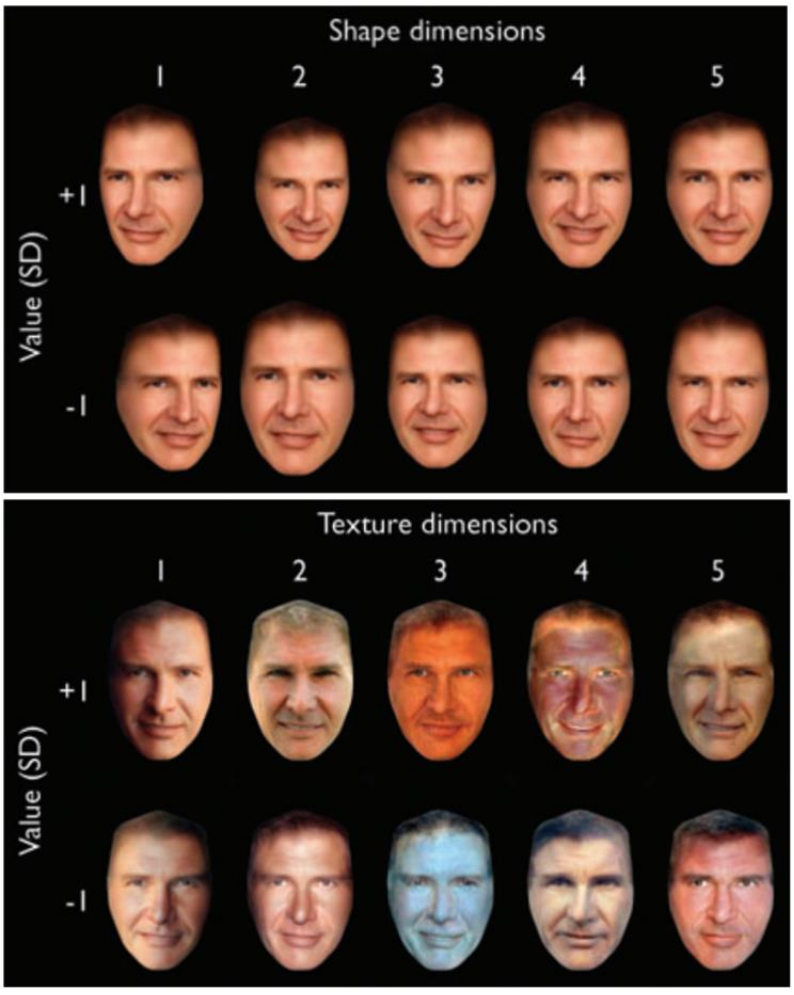

Figure 1.8. Shape (top) and texture (bottom) components derived from 48 images of Harrison Ford (Burton et al., 2011). Each column represents a single component with values *z* = +1 above and -1 below.

1.5. Social Evaluation Models

The first social evaluation model was based on a PCA-trained classifier of 2D computer-generated composite faces, all with a neutral expression and hair/accessories removed (Brahnam, 2005). Images were rated for adjustment, warmth, dominance, trustworthiness, and sociality and then separated into three classes (high, neutral, and low) depending on their ratings. Faces in the neutral category were excluded from further analysis as classification was not unambiguous for those faces. A separate PCA was then trained for each social attribute by randomly dividing the images into a training and a testing set, extracting the eigenvectors from the training set and calculating the distribution of each category within the face space. Test images were first projected onto the face spaces obtained from the training images and the best-fit category membership was then determined. This approach produced good classification rates for all five social traits ranging from 64% (dominance) to 89% (warmth). Branham further used PCA to reconstruct novel images with a high probability of eliciting specific trait attribution. She created two PCA spaces (high and low) for each trait and created new faces by projecting and reconstructing an image from one category (say, the high trustworthiness set) onto the space of the opposite category (low trustworthiness set). These novel images were then rated by human participants and results followed their predicted direction for all five social traits (e.g. faces projected into the low PCA space were rated at the lower end of the scale compared to faces projected into the high PCA space). This was therefore the first successful attempt to use a face recognition approach to classify social, rather than factual (identity or gender) dimensions and to predict social trait attribution.

Developing this PCA approach further Todorov and colleagues developed the most influential data-driven computational model of first impressions (Oosterhof & Todorov, 2008; Todorov, Dotsch, Wigboldus, & Said, 2011). In contrast to Brahnam (2005), Todorov aimed to extract all physical information in the face that is used to inform these social judgements. In order to do so, they firstly collected unconstrained descriptions of face images and extracted 14 dimensions which best accounted for the over 1100 original

traits. The same faces were then rated specifically for these 14 social attributes and the data was subjected to PCA which identified a two-component solution (see Figure 1.9). The first component accounted for 63.3% of the variance and had high positive loadings from all positive traits and negative loadings from all negative traits so it was referred to as valence evaluation. The second component explained 18.35% of the variance and had high loadings from judgements of dominance, aggressiveness, and confidence so it was interpreted as dominance evaluation. Out of all 14 social traits, judgements of trustworthiness corresponded best to the first component and judgements of dominance corresponded best to the second component.

These findings were further tested by performing PCA on different sets of trait judgements, supporting the idea that trustworthiness and dominance can be regarded as reflecting the underlying dimensions of the evaluation of emotionally neutral faces – valence and dominance. This two-component solution is also consistent with other dimensional models, including models of concept evaluation (Osgood, Suci, & Tannenbaum, 1957) and models of interpersonal perception (Wiggins, 1979) which have also been shown to rely on two orthogonal dimensions – affiliation and dominance. Furthermore, the dimensions of trustworthiness and dominance also fit well with the dimensions of warmth and competence which have been identified as fundamental in personality research and perceptions of cultural groups in particular (Cuddy, Fiske, & Glick, 2008).

Oosterhof and Todorov (2008) argue that trustworthiness and dominance are the fundamental underlying dimensions of social face evaluation as they reflect the appraisal of threat. Trustworthiness ratings are related to the perceived intention to help or cause harm and are largely based on emotion generalisation, where faces displaying negative emotions are perceived as untrustworthy, whereas faces displaying positive emotions are perceived as trustworthy (Todorov, 2008). The dominance dimension then reflects the perceived ability to perform these helpful or harmful intentions.

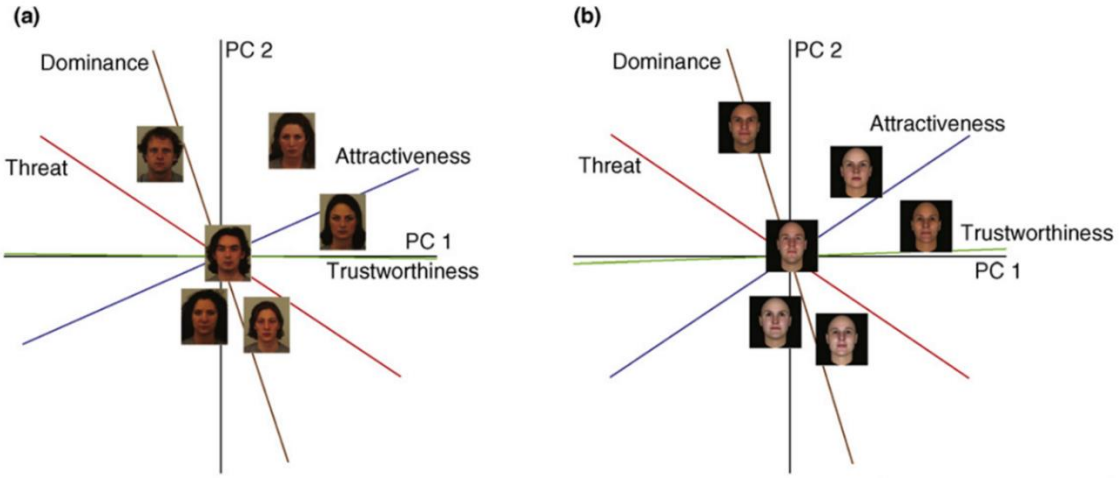

Figure 1.9. The structure of face evaluation as described in Todorov (2008) following the analysis of 66 natural faces (a) and 300 computer-generated faces (b).

Such first impressions models are generally based on two preconditions: 1) that there is high inter-rater reliability and 2) that the model accounts for a meaningful proportion of variance in these judgements, implying that the judgements are systematically associated with certain facial characteristics defined in the face model. In order to model the variability in these two social traits and identify the underlying facial characteristics that govern these judgements, Oosterhof and Todorov used a statistical data-driven PCA approach based on 3D laser scans of face images (the Facegen model). The analysis of the faces followed a procedure similar to Blanz and Vetter (1999) where shape and texture were each represented by 50 dimensions. After establishing high levels of rater agreement (Cronbach's alpha ranging from .76 to .92 depending on social trait) and meaningful proportion of the variance in these judgements explained by the model (ranging from .56 to .91 for shape and texture together), Oosterhof and Todorov constructed new dimensions in the face space that could account for the maximum variability in the judgements. These new dimensions were then used to visualise the changes in the face relative to each social attribute. Figure 1.10 shows social continua of faces created by increasing the weighting of the new dimensions.

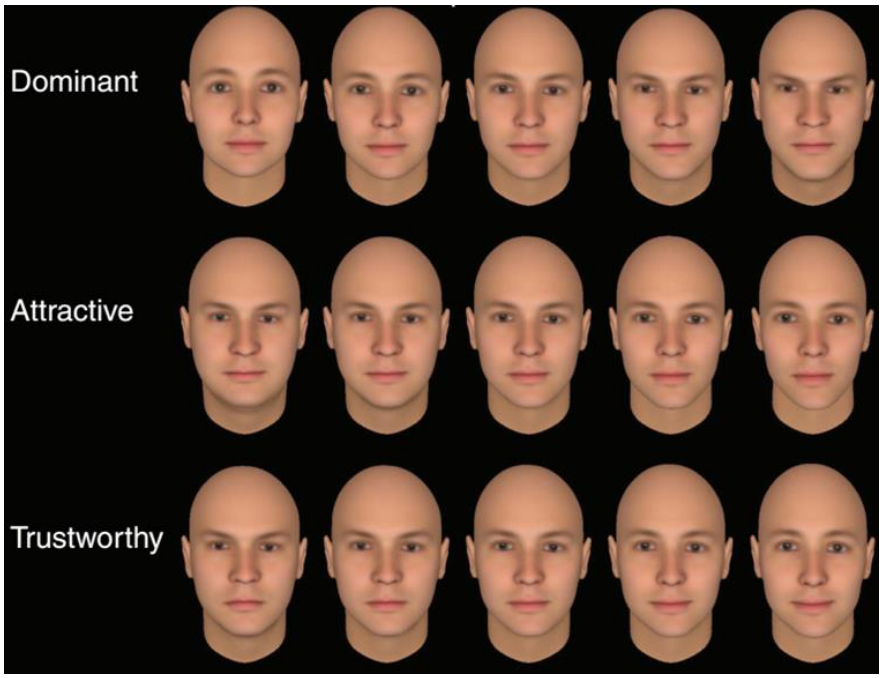

Figure 1.10. Continua of faces demonstrating information in the face relevant to each social trait. The perceived value of the faces on the respective dimensions increases from left to right.

Model validation was obtained by creating seven different variations of the same face that varied on trustworthiness and dominance and collecting ratings of these artificially-created images (Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013). Results showed that both trustworthiness and dominance ratings followed their predicted pattern, demonstrating that the model had successfully captured the underlying information in the face diagnostic of social evaluation. Nevertheless, this model was based on the information gathered from a set of constrained and neutral face images and while this could mean further control over the experimental stimuli, it also means that a large amount of natural face variability was not taken into consideration.

Another face model that aimed to explore the underlying facial characteristics and features people might use to inform their social attribute judgements is the Basel Face Model developed by Walker and Vetter (2009). Just as Todorov, they adopted this approach for personality visualisation in

faces based on the idea that there is high inter-rater reliability in personality judgements, implying that raters use facial information to inform their judgements. To develop their model Walker and Vetter used the facial information from 100 male and 100 female three-dimensional registered laser scans of faces and collected ratings of aggressiveness, attractiveness, extroversion, likeability, risk seeking, social skills, and trustworthiness (see Figure 1.11 for a schematic representation). It is important to note that these faces were displayed in colour, frontal view, had the same lighting and neutral facial expressions. Also, no additional information such as hair, facial hair, or make-up was available.

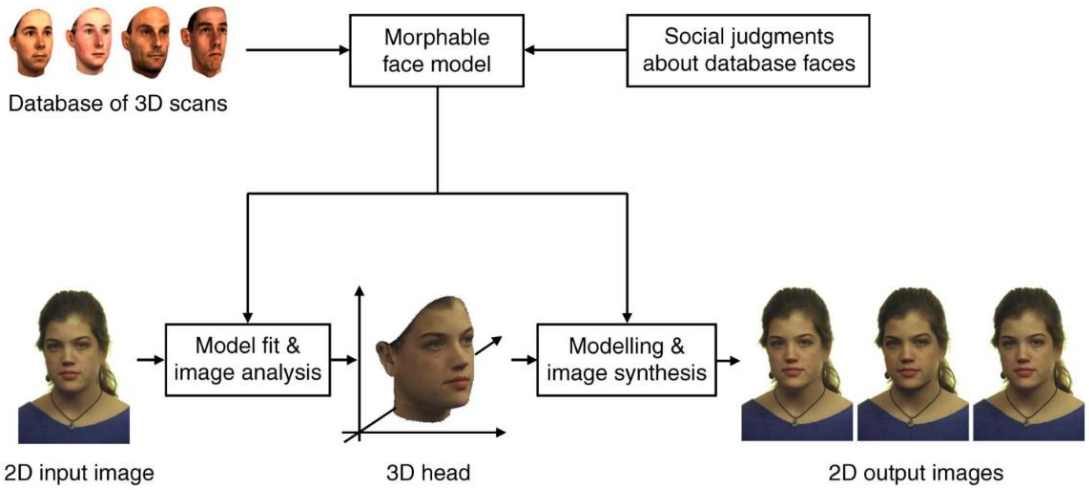

Figure 1.11. Schematic representation of the image analysis for the Basel Face Model (Walker & Vetter, 2009).

PCA on these social attributes identified a two-component solution which explained a total of 77% of the variance. The first component explained 52% of the variance and had high loadings from social skills, likeability, attractiveness, extroversion, and trustworthiness and the second component explained 25% of the variance and had high loadings from risk seeking and aggressiveness. These results are consistent with Todorov's as the two components fit well with the two basic dimensions of face evaluation – trustworthiness and dominance.

Walker and Vetter used a PCA approach similar to that of Oosterhof and Todorov to extract shape and texture information in the faces and then identified the physical dimensions, which captured the variability in ratings of each personality trait. The relative weighting of these dimensions was then manipulated in order to bring about changes in social perception. This was applied both to the faces from the original analysis and completely novel images (Blanz & Vetter, 1999). The validity of this procedure was obtained by creating pairs of face images where the same base face was manipulated slightly towards a higher and slightly towards a lesser degree of the same attribute and these were then used in a 2AFC task (see Figure 1.12 for examples of manipulated images). The results demonstrated that people were able to identify the direction in which all attributes were manipulated above chance levels, with accuracy percentages ranging from 61% (risk seeking) to 100% (likeability), implying that the underlying characteristics responsible for differences in social attribute judgements have been successfully identified and manipulated.

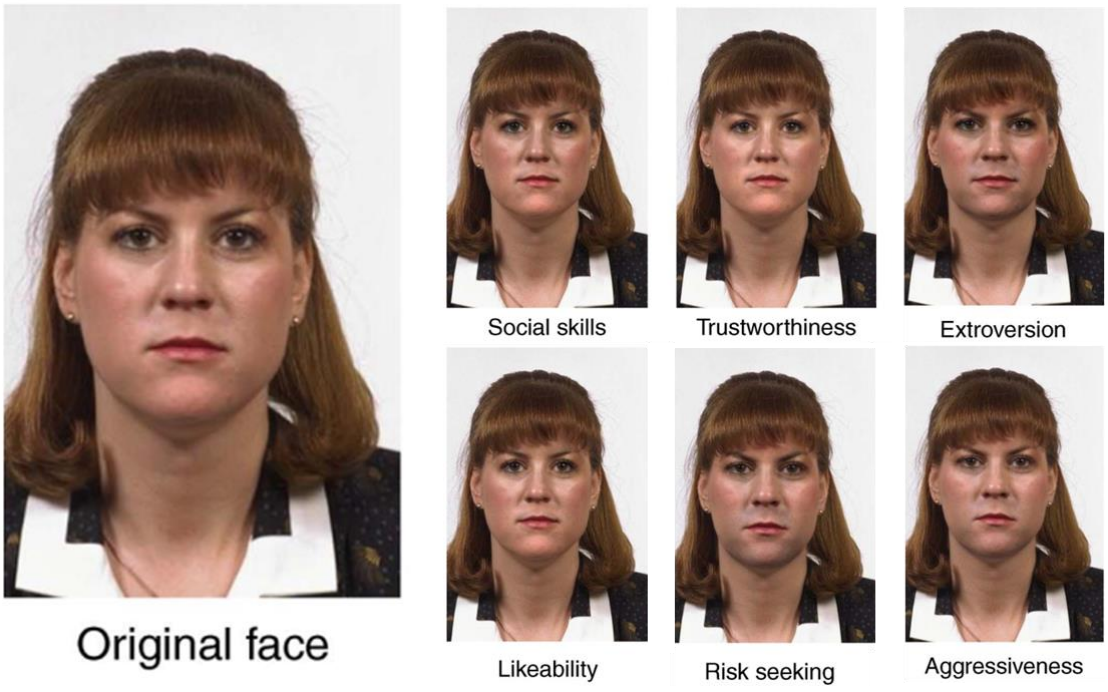

Figure 1.12. Image manipulation using the Basel Face Model. Images on the right are manipulated to represent each social trait to a greater extent.

Walker and Vetter (2015) have also recently extended their model in an attempt to manipulate the perception of the Big Two (communion and agency) and the Big Five (openness, conscientiousness, extraversion, agreeableness, and neuroticism) personality traits. Their results demonstrated not only that the Big Two dimensions fit well with the dimensions for the two basic face evaluation traits – trustworthiness and dominance, but also that it is possible to successfully model and manipulate four out of the Big Five traits.

Reverse correlation models

While all of these face evaluation models undoubtedly capture the underlying information people use to inform their first impressions, it is still difficult to verbalise which features or feature configurations in the face are diagnostic to different social attributes. This is due to the holistic and linear nature of information extracted in PCA-based models as well as the infinitely large number of possible feature combinations that can contribute to social evaluation (e.g. 15 binary features would result in 32,768 possible combinations; Todorov et al., 2011). A data-driven reverse correlation (RC) approach has been suggested as a potential way to address these issues and identify the specific features involved in social evaluation. It was originally developed in auditory perception (Ahumada & Lovell, 1971) and has since been adapted for use in neuropsychology (Ringach & Shapley, 2004) as well as visual (Beard & Ahumada, 1998 ; Solomon, 2002) and social cognitive research (Dotsch, Wigboldus, Langner, & Van Knippenberg, 2008; Karremans, Dotsch, & Corneille, 2011). In face perception, Mangini and Biederman (2004) have used the RC technique to model identities, gender, and emotional expressions.

Recently, RC has also been used to extend existing face evaluation models by increasing the specificity of the extracted information and quantifying diagnostic face regions (Dotsch & Todorov, 2012). Moreover, compared to other models, this approach makes use of both shape and texture information and preserves external features, such as hairstyle, all of which have been shown to be important cues in social evaluation (Macrae & Martin, 2007; Todorov & Oosterhof, 2011).

The process involves creating pairs of images from the same base face by superimposing a randomly-generated noise pattern to one image and the mathematical negative of that same pattern (i.e. pixels that were light in the original pattern became dark in its negative) to the other image. Participants are then presented with both images and asked to identify the image that fits a specific social trait to a greater extent. Finally, all noise patterns selected as representative of these traits are averaged together and superimposed on the base image to create a classification image (CI), while their unselected counterparts are averaged and superimposed to produce an anti-CI. Using this technique, Dotsch and Todorov (2012) showed that features diagnostic of trustworthiness include a smile, wide open almond-shaped eyes, and a smooth small face, whereas thicker downturned lips, angry-looking eyes, a square-shaped face, and sagging cheeks are the features diagnostic of untrustworthy faces (see Figure 1.13 for examples of CI and anti-CI images). Dominance, on the other hand, is associated with strong eyebrows, slightly downturned mouth, and dark narrow eyes, while submissive faces have sad-looking eyes and thin frowning lips. Also, a greater contrast between face and background signals dominance, whereas the submissive face has a less distinctive outline. This model was further validated by asking a new sample of participants to rate the CIs and anti-CIs for dominance and trustworthiness. Ratings followed the predicted direction, demonstrating that this RC approach had successfully captured features diagnostic of the fundamental social evaluation dimensions.

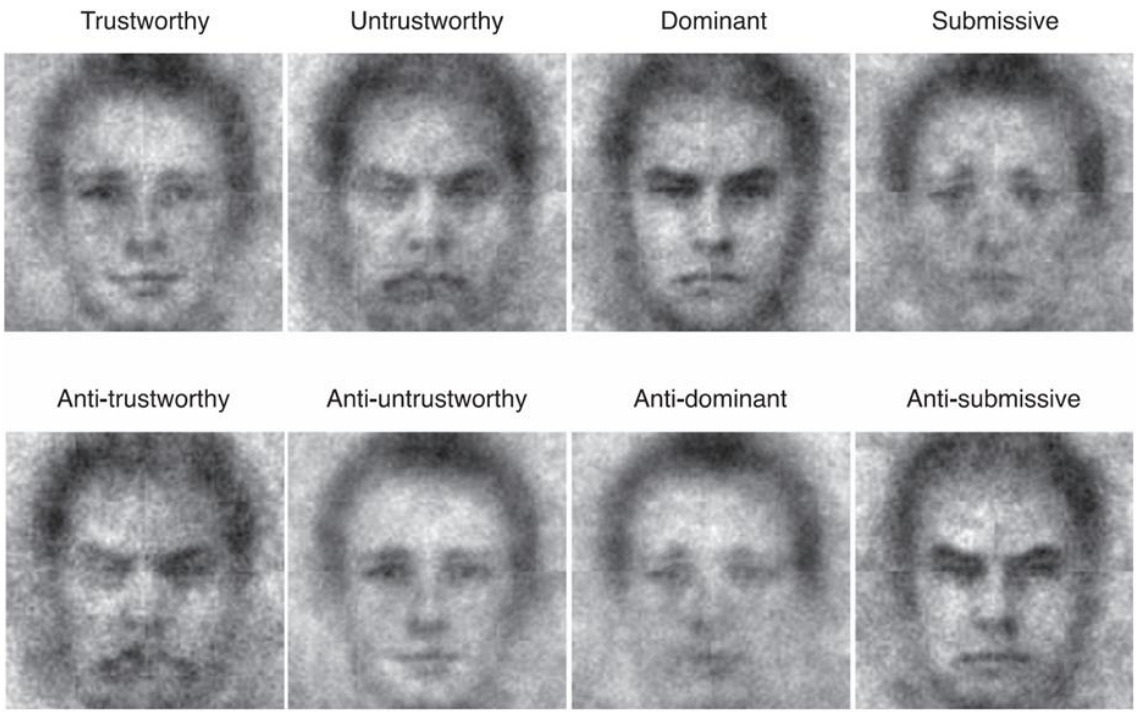

Figure 1.13. Classification (top row) and anti-classification (bottom row) images from Dotsch and Todorov (2012). Classification images are the average of all noise patterns selected as best resembling the target social trait, superimposed on the base image, while anti-classification images are the result of patterns not selected as resembling the target trait, superimposed on the base image.

All of the models described so far capture the underlying information in the face diagnostic of social evaluation. Using this information to manipulate or validate perception, however, is usually associated with changes in face shape and consequently – identity. Looking at Figure 9, for example, there is a clear change in identity between the two ends of social trait spectra. Robinson, Blais, Duncan, Forget, and Fiset (2014) therefore suggested a different RC technique that can be used to manipulate social perception without much alteration of facial features. This is accomplished with the Bubbles method (Gosselin & Schyns, 2001, see Gosselin & Schyns, 2004 for a comparison between RC and Bubbles) where Gaussian "bubbles" are applied to faces in order to reveal trait-diagnostic feature configurations. To produce their Bubbilised faces, Robinson et al. filtered face images into five

non-overlapping spatial frequency bands and sampled each band using Gaussian windows of varying standard deviations (see Figure 1.14). These five images were then combined to create the final stimulus set, which, together with the original images, was rated for trustworthiness and dominance.

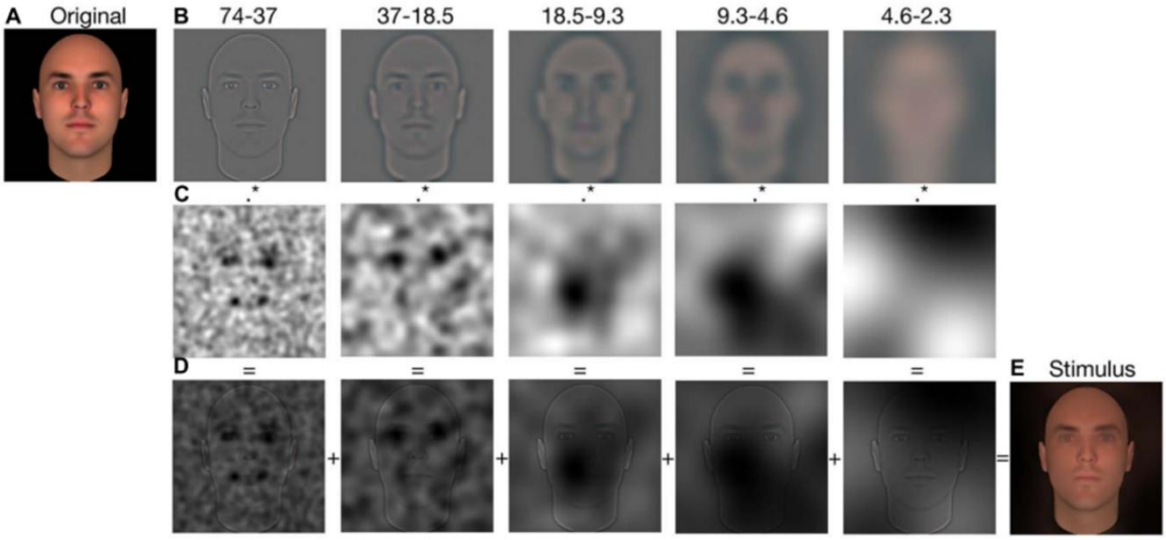

Figure 1.14. Stimuli generation used in Robinson et al. (2014). Each original image (A) was firstly decomposed into five spatial-frequency bandwidths (B). Each bandwidth was then multiplied by the respective classification image (C) and the resulting information was summed across the five scales (D) to produce the filtered stimulus (E).

Results revealed the eyes and mouth as diagnostic areas for trustworthiness. Changes in the eyes and eyebrows led to increased perception of dominance, whereas the mouth and jaw were mostly diagnostic of perceived submissiveness. Revealing or hiding this trait-diagnostic information in any face was then shown to manipulate social evaluation using both computer-generated and real face images (see Figure 1.15 for an example), validating the approach. Such findings demonstrate that social evaluation can be changed within a single identity which is particularly relevant to the work described here.

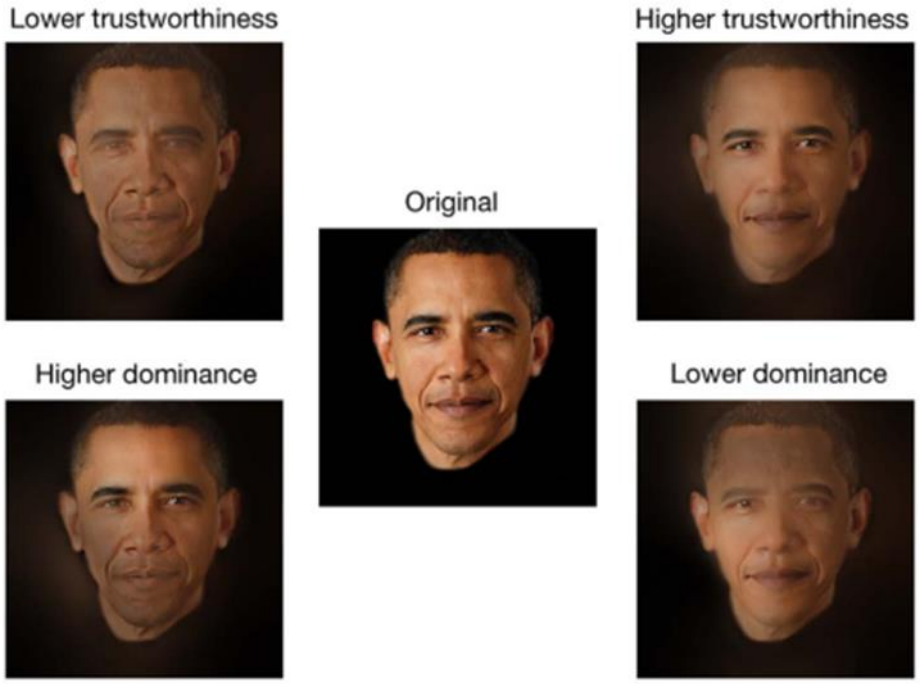

Figure 1.15. Manipulating social perception with information extracted using the Bubbles technique (Robinson et al., 2014).

Taken together these studies show that regardless of how subjective or arbitrary social attributes may seem, the high levels of agreement in social judgement attribution across people enables us to identify the physical dimensions we use to inform our judgements. It is important to note, however, that the manipulation of social attributes here is independent of identity as the physical analysis was performed on images of different people. This therefore implies that changing the perception of a certain social attribute could also change that person's identity as well.

1.6. Within-Person Variability

The face models developed by Oosterhof and Todorov (2008) and Walker and Vetter (2009) were very successful and of importance here is that the image sets used to inform these models were tightly controlled ones. Although this allows a more precise analysis and manipulation of facial parameters, it inevitably fails to incorporate and account for the considerable face variability

in real faces and images of faces encountered in everyday life. To this end Jenkins, White, Van Montfort, and Burton (2011) have recently introduced the concept of 'ambient images' and demonstrated that the processing of this natural face variability can also be used to get a better grasp on identity recognition and within-person variability. Ambient images are highly variable and much more representative of the diverse conditions under which we see faces in our everyday lives. They encompass differences in the person, such as the range of emotional expressions, aging or facial hair and make-up as well as differences in the world, such as angle of photograph and lighting direction. The importance and meaningfulness of sampling this natural face variability was highlighted by Sutherland et al. (2013) who collected ratings for a large database of 1000 ambient images on 13 social attributes, including trustworthiness, approachability, dominance, confidence, etc. Using PCA on these ratings identified a third factor, youthful-attractiveness, that emerged in addition to the main social evaluation traits – trustworthiness and dominance (see Figure 1.16). This highlights how dependent each face evaluation model is on the specific images used to inform it and demonstrates that with real world variability, social evaluation might have a more elaborate underlying structure than it was previously thought.

Another vastly underestimated aspect of face perception, closely related to the concept of natural ambient images, is within-person variability. Face perception research so far has mostly focused on between-person variability or how to tell people apart, while ignoring the importance of how to tell people together. Nevertheless, a few studies on the face familiarity effect have recently tried to address this gap and highlight how valuable it is to gain a deeper understanding of within- as well as between-person variability (Burton et al., 2011). Further support comes from a card sorting task where participants are presented with 40 face images and instructed to sort them by identity in a way that images of the same person are grouped together (Jenkins et al., 2011). These were in fact 20 images of two Dutch celebrities who were, critically, unfamiliar to UK viewers. Surprisingly, UK participants sorted these images into nine identities on average while Dutch participants

who were familiar with these identities sorted the cards almost perfectly in two piles. These findings not only highlight the effect of familiarity in face perception but also demonstrate that difficulties in telling people together or coping with within-person variability plays a significant role in face recognition and perception errors. Extending these findings to the computational and automatic face recognition context, variability in faces has been utilised to generalise knowledge to novel face exemplars, including modelling of changes in lighting or viewpoint (O'Toole, Edelman, & Bulthoff, 1998). Moreover, automatic face recognition approaches based on within-person PCA (or face-specific subspace PCA as it is referred to in the literature) have been shown to improve recognition accuracy (Aishwarya & Marcus, 2010; Shan, Gao, & Zhao, 2003).

Figure 1.16. Manipulation of social perception using information extracted from 'ambient' images (Sutherland et al., 2013).

Most importantly, a recent key paper by Todorov and Porter (2014) highlights the importance of within-person variability in social evaluation. They collected ratings for five different images of the same identity and compared rating variability in images of different identities with rating variability in images of the same identity. Results showed that within-person variability exceeded or was at least comparable to between-person variability for the attribution of trustworthiness, extraversion, meanness and creativity. Between-person variability was much larger for judgements of attractiveness only, suggesting that attractiveness might be linked to identity to a greater extent than other social traits. Moreover, the relative order of identities according to trait ratings could easily be reversed by sampling different images of the same identities. As can be seen in Figure 1.17 differences in identity could be overwritten by image selection.

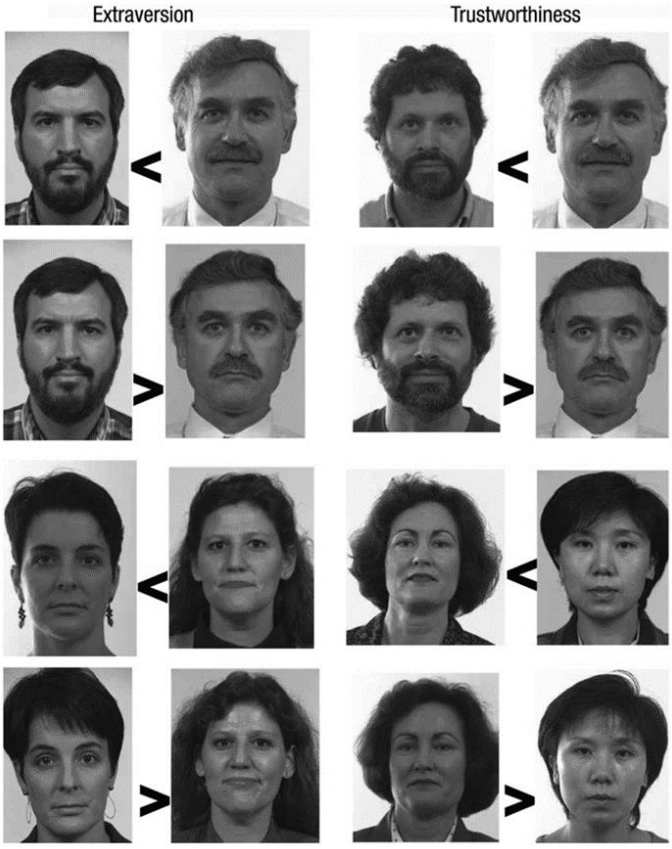

Figure 1.17. Pairs of images demonstrating reversals of attribute ratings of extraversion (left) and trustworthiness (right). For each pair the top row shows images where the person on the right received a higher rating and the bottom row shows images where this relative order is reversed (Todorov & Porter, 2014).

These findings are further supported by Jenkins et al. (2011) who used 20 different images for each of 20 Dutch celebrities and asked participants to make a Yes/No attractiveness judgement, following which each image received an attractiveness score out of 20. As shown in Figure 1.18, there was a great amount of variability in attractiveness scores for images of the same identity, implying that social attribute judgements are not only dependent on identity, but also on within-person variability and the specific image used.

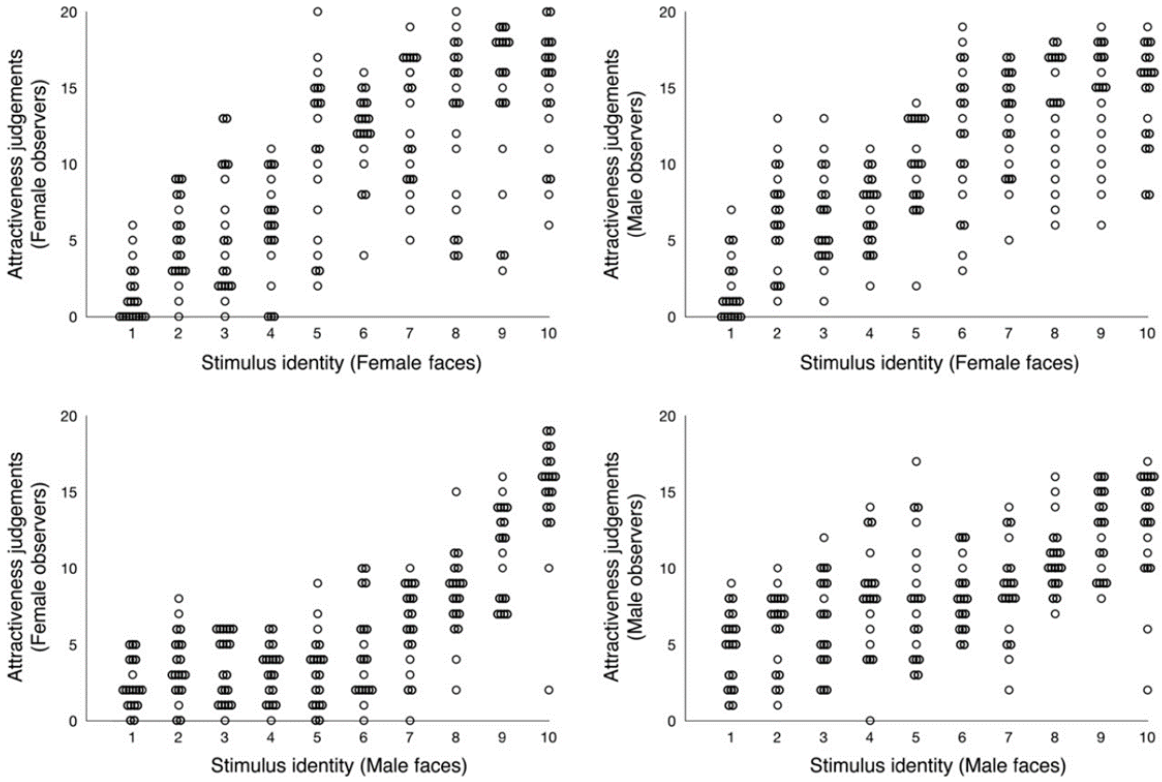

Figure 1.18. Spread of within- and between-person variability in attractiveness scores from Jenkins et al. (2011). Data is shown separately for male (right) and female (left) raters and male (bottom) and female (top) faces. Each column represents a single identity and each point – a single image. Identities are ranked by overall attractiveness.

Although these studies inevitably support the idea of within-person variability in social attribution they are still limited, as the former uses only five different exemplar images per identity which seems insufficient and the

latter only investigates within-person variability in attractiveness ratings. Therefore, there is a lot more to be explored in within-person variability and getting a good grasp of this variability is essential to fully understand face perception and recognition. We are constantly changing (through aging or expressing different emotions) and so is the world around us (reflected in changes in lighting or camera angle), therefore representing a person as a single point in space is a very limited perspective.

1.7. First Impressions Across Modality

A wealth of biological and social information about people, such as sex, age, ethnicity, and emotional state, can be inferred not only from people's faces (Bruce & Young, 1986) but from their voices as well (Belin, Bestelmeyer, Latinus, & Watson, 2011; Yovel & Belin, 2013). This is true for both person recognition, where people can be accurately identified from their faces and voices, and social evaluation, where stable first impressions about unfamiliar others are formed from both facial and vocal cues (Todorov et al., 2009; Zuckerman & Driver, 1989). Some even argue that given their primitive origins, vocal cues may have a more important role in social perception than either linguistic content or other non-verbal cues including, for example, facial characteristics or expressions (Tusing & Dillard, 2000). Social impressions from those two modalities present a number of similarities, encompassing rater consensus, implications, and underlying structure (Ballew & Todorov, 2007; Chen, Halberstam, & Yu, 2016; McAleer, Todorov, & Belin, 2014; Tigue, Borak, O'Connor, Schandl, & Feinberg, 2012; Wilson & Rule, 2016; Zebrowitz & Montepare, 2008). They, however, differ in processing time as first impressions from faces arise very quickly (after less than a second of exposure in many reports), whereas auditory impressions require longer exposure in order for the voice to fully unfold.

First impressions from voices

First evidence that listeners readily form impressions of unknown speakers based on their vocal characteristics come from Pear (1931) and Allport and Cantril (1934) who used radio broadcasts to obtain listener

evaluations of presenters' voices. Since then, these vocal evaluations have been characterised with high level of agreement between listeners demonstrating that people use consistent acoustic information in the voice to inform their social judgements (Zebrowitz & Montepare, 2008). Similarly to face evaluation, impressions from voices have been shown to fall along two fundamental dimensions. Adopting the same approach as Oosterhof and Todorov (2008), McAleer et al. (2014) collected ratings for a number of social traits from brief voices and used PCA to show a two-dimensional social evaluation space, with valence and dominance as the main dimensions. Such findings are consistent not only with models of face evaluation, but also with models of group and concept evaluation (Fiske, Cuddy, & Glick, 2007; Osgood et al., 1957).

Moreover, these zero-acquaintance impressions from voices have been associated with similar social outcomes as the ones from unfamiliar faces. Voting behaviour, for example, can be influenced by both facial and vocal information. While this is linked to perceptions of competence in face evaluation (Ballew & Todorov, 2007), people with low-pitched voices are attributed superior leadership abilities and consequently receive more votes (Klofstad, Anderson, & Peters, 2012; Tigue et al., 2012). This can further be related to perceptions of cooperativeness where feminine (i.e. high) pitch qualities are associated with higher levels of friendliness and collaboration and lower likelihood of threatening behaviours (Knowles & Little, 2016). Just as with faces, certain voice features have been shown to predict mating and reproductive success (Hodges-Simeon, Gaulin, & Puts, 2011; Puts, 2005). Recent studies have even shown that ratings of men's vocal attractiveness by unfamiliar women can predict the same ratings made by women familiar with those identities (Doll et al., 2014), implying some degree of accuracy in those attributions. The pervasive effects of vocal first impressions have also been demonstrated in the courtroom, a situation where decisions should be based on objective evidence rather than subjective impressions. Chen et al. (2016), for example, asked participants to rate lawyers' opening statements before the Supreme Court between 1998 and 2012 for a number of social traits including masculinity, attractiveness, trustworthiness, and aggressiveness

and reported a higher proportion of won cases for voices rated as less masculine.

Voice pitch (also referred to as mean fundamental frequency, F0) is one of the most perceptually salient acoustic cues used by listeners to infer emotion and affect in speech (Dimos, Dick, & Dellwo, 2015). It is also used to evaluate voices on socially-relevant dimensions including attractiveness, trustworthiness, and dominance (Feinberg et al., 2006; Jones, Feinberg, DeBruine, Little, & Vukovic, 2010; Puts, Hodges, Cárdenas, & Gaulin, 2007; Wolff & Puts, 2010). F0 is determined by vocal fold length and the tension applied to those folds, and is a sexually dimorphic characteristic, with adult males speaking at a lower vocal pitch than adult females on average (Titze & Martin, 1998). Perceptions of attractiveness are affected by voice quality measures related to both pitch (deepness, squeakiness and throatiness) and impact (monotonous, loudness and resonance; Zuckerman & Miyake, 1993) in a directional way across male and female identities. That is, male participants evaluate high-pitched female voices as more attractive, whereas female participants show a preference for low-pitched male voices (Bruckert, Liénard, Lacroix, Kreutzer, & Leboucher, 2006; Collins & Missing, 2003; Feinberg, Jones, Little, Burt, & Perrett, 2005). These differences have been explained in terms of evolutionary theories arguing that lower pitch in males is a signal of good fitness and higher levels of testosterone (Feinberg et al. 2005; Harries, Walker, Williams, Hawkins, & Hughes, 1997; Hollien, Green, & Massey 1994; Puts, 2005) as well as through female sensitivity bias for processing low-pitched sounds (Hunter, Phang, Lee, & Woodruff, 2005). There is even evidence that low voice pitch is related to reproductive success in an indigenous tribe of hunter-gatherers (Apicella, Feinberg, & Marlowe, 2007).

The link between vocal pitch and listener perceptions of dominance is particularly well researched. Building on work by Morton (1977), who argued that lowered pitch marks aggression and dominance across a variety of animal species, Ohala (1984) showed that low-pitched human recordings were rated as sounding more dominant than high-pitched recordings when all other voice aspects remained constant. He further argued that the lowering of

mean pitch to signal dominance is related to body size, described by the 'frequency code' or 'size code' hypothesis (Ohala, Hinton, & Nichols, 1997). This argument, however, is centred around speakers' attempts to manipulate their pitch in order to *appear* more physically dominant and later studies report no statistically significant relationship between speaker height, weight and F0 in either running speech (Kunzel, 1989) or single vowels (Gonzalez, 2004). This highlights the lack of one-to-one mapping between pitch and body size, and suggests that the relationship between pitch and dominance is more closely tied to listeners' perceptions of speaker size than it is in the biological relationship between body size and F0. Indeed, a range of perceptual studies have shown a link between perceptions of body size, personality judgements, and the lowering or raising of F0 (Chuenwattanapranithi, Thipakorn, & Maneewongvatana, 2009; Xu & Kelly, 2010).

In terms of social evaluation, low-pitched male voices have been consistently associated with hostility and aggressiveness, whereas a higher pitch is usually perceived as submissive and non-threatening (Tusing & Dillard, 2000; Vucovic et al., 2011). Puts, Gaulin, and Verdolini (2006) reported that a single semitone increase or decrease in mean pitch caused listeners to perceive significant differences in both social and physical dominance for male speakers, with lowered pitch resulting in higher dominance ratings. After establishing that voice evaluation follows the same two-dimensional structure as face evaluation, McAleer et al. (2014) explored acoustical measures that could potentially account for the variance in those two dimensions (trustworthiness and dominance). Their results showed that alpha, F0, harmonic-to-noise ratio (HNR, roughness of voice) and formant dispersion (measured as the ratio between consecutive formant means) explained 68% of the variance in male voices, whereas only F0 and dispersion predicted dominance ratings for female faces (explained only 27% of variance). Critically, higher dominance was related to lower pitch in male voices but higher pitch in female voices. Therefore, while pitch seems to be an important cue to dominance across both genders, it acts differently for male and female voices and its effect is generally more pronounced when evaluating male vocal recordings. Some contradictory findings also highlight

that the perceptual link between dominance and lowered pitch is less well-established for female voices than for male voices. Borkowska and Pawlowski (2011), for example, reported higher dominance ratings for lower-pitched female voices and Tsantani, Belin, Paterson, and McAleer (2016) found that while both lower-pitched male and female voices were picked more frequently as the dominant-sounding voice by listeners in a forced choice task, the preference was only significantly greater than chance for the male voices.

Reports of the relationship between vocal pitch and perceptions of trustworthiness are also inconsistent and contradictory with some studies showing a positive relationship (McAleer et al., 2014), some – a negative relationship (low pitch perceived as more trustworthy; Apple, Streeter, & Krauss, 1979; Tsantani et al., 2016), and others no relationship at all (Vukovic et al., 2011). McAleer et al. (2014) further found that F0 and HNR explain 49% of variance in trustworthiness ratings of male voices, whereas pitch had no influence on the perception of trustworthiness in female voices, demonstrating that the relationship between pitch and trustworthiness might be moderated by gender. It is possible that such divergent findings result from the differences in the context and vocal stimuli used in those studies. Vukovic et al. (2011), for example, explores dating and relationship preferences, whereas Klofstad et al. (2012) and Tigue et al. (2012) focused on the election of political representatives. Moreover, studies either use contentful utterances (McAleer et al., 2014) or socially irrelevant vowel sounds (Jones et al., 2010). Therefore, with research suggesting that pitch preferences may vary with social context (Jones, Feinberg, DeBruine, Little, & Vukovic, 2008; Vukovic et al., 2008), it can be argued that these differences are the source of inconsistencies in trustworthiness perception. The relationship between pitch and dominance, on the other hand, is further strengthened by their sexually-dimorphic nature which makes it a more consistent and reliable finding.

Audiovisual integration in social evaluation

Audiovisual integration has been shown to manifest both facilitation and interference effects: Common facilitative influences have been

demonstrated in speech intelligibility or 'lip-reading' where presenting participants with visual information from a speaker's face can significantly improve speech content recognition (from 23% to 65% in Summerfield, 1979). This is also seen in priming studies where participants are quicker to identify a face as familiar after being presented with the voice of that same identity and vice versa (Ellis, Jones, & Mosdell, 1997; Schweinberger, Herholz, & Stief, 1997). A classic interference effect comes from the McGurk illusion (McGurk & MacDonald, 1976) in speech perception where participants are presented with incongruent audio and visual cues and yet integrate them together. Attending to a video clip of a person pronouncing the syllable /ga/ while listening to a superimposed audio clip of a person pronouncing the syllable /ba/, for example, results in the impression that the person in the video clip actually pronounces the syllable /da/.

While both voices and faces provide us with a wealth of social information (Bruce & Young, 1986; Belin et al., 2011) and there is a multitude of studies investigating the independent effects of facial and vocal cues on social perception (Berry, 1990; Hodges-Simeon, Gaulin, & Puts, 2010; Oosterhof & Todorov, 2008; Zuckermann & Driver, 1989), existing audio-visual integration research has been almost exclusively focused on emotion and identity recognition (see Campanella & Belin, 2007 for a review). Massaro and Egen (1996), for example, presented participants with congruent and incongruent face-voice pairings where face images displayed happy, angry, or neutral expressions, while the voice stimuli were created by an actor pronouncing the word *"please"* in a happy, angry, or neutral way. Participants' task was to simply classify the emotion as happy or angry. The study showed that while both facial and vocal cues were effective for expression categorisation, visual information from the face had a stronger effect as it changed performance across all three voice emotion levels. Such results are consistent with the general finding that faces seem to be more reliable cues than voices in emotion recognition (Hess, Kappas & Scherer, 1988; Mehrabian & Ferris, 1967).

Audio and visual signals are informative not only with respect to emotion classification but also to identity recognition. The effect of face and voice cues seems to be additive as studies report an intermediate speed of familiarity judgements to previously learned face-voice associations bimodally compared to unimodal face (fastest) and voice (slowest) conditions (Joassin, Maurage, Bruyer, Crommelinck, & Campanella, 2004). Furthermore, similarly to integration in emotion perception, the visual channel has been shown to be of greater importance for identity decisions (Ellis et al., 1997; Schweinberger et al., 1997). In a series of familiarity decision studies Schweinberger (Schweinberger, Kloth, & Robertson, 2011; Schweinberger, Robertson, & Kaufmann, 2007) provided evidence for audio-visual integration in person identification across modalities (face and voice together, voice only, face only), congruency (faces and voices combined were of the same identity or not), and levels of realism and synchronicity (voices paired with either static or time-synchronised articulating faces). Results showed best performance in terms of both reaction time and accuracy for static and dynamic familiar faces paired with the voice of the same identity, followed by unimodal voice only presentation and worst performance for static and dynamic faces paired with a non-corresponding voice. Authors also demonstrated that while both static and dynamic corresponding faces brought about significant facilitation in response time compared to baseline (voice only presentation), this effect was much more pronounced for dynamic faces, highlighting the importance of synchronicity and a more realistic audio-visual integration.

Much less is known about audiovisual integration in social perception with only a few studies pairing faces and voices together and collecting social attribute ratings. This is surprising considering the great many studies on the independent effects of facial and vocal cues on social evaluation (Oosterhof & Todorov, 2008; Zuckermann & Driver, 1989). Based on the 'halo effect' or the idea that physically attractive people are evaluated more favourably on other socially-relevant dimensions, Zuckerman (e.g., Zuckerman, Hodgins, & Miyake, 1990; Zuckerman, Miyake, & Hodgins, 1991; Zuckerman, Miyake, & Elkin, 1995) investigated the relative effects of facial and vocal attractiveness on trait judgements using congruent (e.g. both face and voice high in

attractiveness) and incongruent (e.g. a face high in attractiveness paired with a low-attractiveness voice) face-voice pairings. Consistent across a series of studies, results showed that both face and voice attractiveness contributed significantly to subsequent social attribute ratings. There was also a significant interaction between the two, suggesting a synergistic model, where social ratings of high-physical and high-vocal attractiveness pairing were relatively high and ratings of all other pairings were relatively low. Focusing on the relative and combined effects of face and voice cues, rather than their attractiveness specifically, Rezlescu et al. (2015) collected trustworthiness, dominance, and attractiveness ratings of faces and voices, uni- and bi-modally. Their findings showed different weighting of audio and visual information depending on social trait. Visual information from the face was more salient for judgements of attractiveness, whereas dominance attribution relied mostly on audio information, although it should be noted that main effects of both face and voice were significant for the evaluation of those traits. Judgements of trustworthiness, on the other hand, were based on face and voice cues equally as well as on their combination, demonstrated by a significant face-voice interaction. Other recent studies, however, challenge these findings, reporting a greater effect of facial cues over vocal ones for trustworthiness attribution (Tsankova et al., 2015). This might be due to the type of vocal stimuli paired with the faces. Rezlescu recorded participants pronouncing English vowels with a neutral expression in order to control for the potential effects of emotional prosody, accent, or inflection, whereas Tsankova used contentful utterances, arguing this is more representative of real life first impression situations.

There is evidence for the automatic nature of audiovisual integration in terms of both emotion categorisation and identity recognition with studies showing that this cross-modal effect occurs in the absence of conscious perception. De Gelder and Vroomen (2000), for example, presented participants with face images making up a continuum between a happy and sad expression paired with vocal recordings of content-neutral sentences pronounced in a happy or sad way. Participants' task was to classify the emotion of each identity (face and voice presented simultaneously) based on

either auditory or visual cues only. Results revealed that participants were able to follow experimental instructions as demonstrated by the reduced cross-modal bias in this focused attention paradigm compared to decisions based on both facial and vocal cues. Nevertheless, the channel that was meant to be ignored still had a significant effect on emotion classification, implying an automatic and mandatory integration process. This is further supported by identity recognition studies, where participants are presented with corresponding and non-corresponding face-voice pairs and required to make a familiarity judgement (Schweinberger et al., 2007, 2011). Recognition of a familiar voice was shown to be both faster and more accurate when paired with the face of the same identity even when participants were specifically instructed to base their decisions on the audio cues only. While the mandatory nature of audio-visual integration is supported for emotion and identity processing, however, no study has explored its automaticity in social evaluation.

1.8. Aims and Overview

This thesis aims to establish the spread and magnitude of within-person variability in social evaluation and use it to extend previous first impression models. In addition to idiosyncratic variability, experiments described here aim to explore social evaluation across gender, familiarity and modality as well as investigate the implications of social attribution for face recognition.

The first experimental chapter (Chapter 2) explores the effect of established face perception and recognition factors on first impressions. In particular, the experiments in this chapter focus on social evaluation and the relationships among social traits across gender and familiarity. The chapter further introduces the process of image averaging used primarily in the face *recognition* literature and explores the *social* information conveyed by such images.

Chapter 3 describes the first attempt at incorporating within-person variability in models of social evaluation. All previous studies have adopted a between-person approach by representing an identity with a single image

(Oosterhof & Todorov, 2008; Walker & Vetter, 2009). Following from studies showing different images of the same person lead to different social attribute ratings (Jenkins et al., 2011; Todorov & Porter, 2014), experiments in this chapter use multiple images per identity that vary in lighting, pose, emotional expression, etc. and therefore represent real-life social interactions more accurately. Thus, these experiments aim to explore the spread and magnitude of within- and between-person variability in social evaluation. They first identify the information in the face people use to inform their first impressions by extracting both shared (between-person) and idiosyncratic (within-person) variability together and then establishing whether this could be achieved with idiosyncratic variability only. Finally, the experiments aim to use the extracted information to manipulate social perception without changing identity, which has not been achieved by current social evaluation models.

Experiments in Chapter 4 build on evidence for the influence of social traits on face memory tasks and the automaticity of social evaluation. They explore the relationship between face recognition and social evaluation by establishing the effect of social evaluation on face matching. Chapter 4 further focuses on the main principle of social evaluation – emotion overgeneralisation, and investigates the effect of emotion expression, and smiling in particular, on face matching performance.

Finally, experiments in Chapter 5 investigate multi-modal social evaluation and explore audio-visual integration in the perception of the two fundamental social evaluation dimensions – trustworthiness and dominance. Building on audio-visual studies in person recognition and emotion classification, they go on to establish whether this integration is automatic and mandatory. Critically, experiments in this chapter manipulate visual and vocal stimuli within-person by sampling many different images of the same person and manipulating the pitch of their voice. Thus, this last chapter brings us closer to understanding integrated person perception.

Altogether, experiments in this thesis demonstrate the overarching effect of within-person variability on face perception and recognition. Findings reported here suggest that first impressions depend on both identity and the statistical properties of images and argue that first impression models should be able to account for both sources of variability in order to represent social evaluation fully and more accurately.

Chapter 2 – First Impressions across Gender and Familiarity

2.1. Introduction

Faces are one of the most prevalent and information-rich stimuli in our everyday lives and even though we are always 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). Most importantly, these evaluations affect our choices and behaviours not only in situations where appearance might be relevant (Olivola et al., 2014), but also in situations, where we should be guided by more objective cues, such as court decisions (Eberhardt et al., 2006; Wilson & Rule, 2016) and political elections (Olivola & Todorov, 2010a). While evidence for the accuracy of these personality evaluations is limited (Todorov et al., 2015), people seem to agree on their judgements, implying that they are using some physical information in the face to inform their impressions (Zebrowitz & Montepare, 2008). Taking this assumption one step further, most influential face evaluation models were developed to extract this information and manipulate people's perceptions. Oosterhof and Todorov (2008), for example, gathered ratings of faces for a range of social traits and identified two underlying dimensions in social evaluation – trustworthiness and dominance. Moreover, recent studies making use of 'ambient' images that vary in emotional expression, pose, lighting or camera angle (Jenkins et al., 2011) extracted an additional evaluation dimension – 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 et al., 1957) and interpersonal perception (Wiggins, 1979), little is known about the relationship between these traits and even less about their relationship across different genders.

Gender in social evaluation

The importance of gender for social evaluation can be supported with studies from the social stereotypes literature (Imhoff, Woelki, Hanke, &

Dotsch, 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 inform one's social judgements. Given the similarities between face and general social evaluation models as well as evidence from gender stereotype studies, it is possible that the fundamental social dimensions – trustworthiness and dominance – are attributed differently for male and female faces. It is important to note, however, that while in their face evaluation model Oosterhof and Todorov (2008) argue that the first dimension, trustworthiness, relates to valence and the second dimension, dominance, represents the perception of femininity-masculinity, general models of social evaluation associate the valence dimension with femininity and the dominance dimension with masculinity (Abele & Wojciszke, 2007; Cuddy et al., 2008).

Further direct evidence for the differential effect of gender on social face evaluation comes from Sutherland et al. (2015) who collected ratings for male and female 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 one hand, ratings of masculine-looking female faces showed that they were perceived as significantly more dominant and significantly less attractive and trustworthy, implying 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, ratings of masculine-looking males showed that they were perceived as significantly more dominant and significantly less attractive, 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 trustworthiness were rated as significantly less dominant than female faces low in trustworthiness, whereas no such difference was found between high- and low-dominance male images.

Familiarity

Another factor that has been somewhat overlooked in social evaluation studies is familiarity. Its effect is key for face recognition as demonstrated by face matching studies where participants are presented with a pair of images on the screen and asked to decide whether they are of the same identity or of two different identities (Bruce et al., 2001; Megreya & Burton, 2008). Accuracy on such tasks with unfamiliar faces has been repeatedly shown to be low (Burton et al., 2010; Megreya & Burton, 2006), whereas images of familiar identities can be matched with surprisingly high levels of accuracy, even when these images are heavily degraded (Burton, Wilson, Cowan, & Bruce, 1999). The influence of familiarity has also been shown to overcome within-person variability, which describes the idea that images of the same person can be just as diverse as images of different people. Using a card-sorting task, where participants were provided with a number of face images and asked to sort them by identity, Jenkins et al. (2011) demonstrated that viewers unfamiliar with the identities in the images sorted them into nine piles on average, whereas familiar viewers found within-person variability much easier to incorporate and accurately divided the images into two piles.

At first glance, familiarity effects may seem irrelevant to social face evaluation as first impressions represent zero-acquaintance judgements. However, investigating social attribution and the relationships between the evaluation dimensions for familiar and unfamiliar faces could provide a more direct comparison, allowing us to identify any changes in social attribution brought about by the familiarisation process. Associations between familiarity and certain social attributes follow from the familiar face overgeneralisation (FFO) hypothesis which argues that the benefits of distinguishing friends from strangers have reinforced a tendency to evaluate people who resemble known identities closely in a more favourable way (Zebrowitz & Collins, 1997; Zebrowitz & Montepare, 2008). The FFO is based on the mere exposure effect which demonstrates that previously seen stimuli generally trigger more positive reactions (Zajonc, 1968). It is further supported by evidence of significantly higher attractiveness ratings for previously seen faces (Rhodes et al., 2001, 2005) as well as evidence for the positive relationship between

familiarity and attractiveness ratings (Peskin & Newell, 2004). Incorporating within-person variability in social face evaluation will then allow us to explore the extent of the familiarity effect. As familiarity has been shown to overcome within-person variability in face recognition (Jenkins, et al., 2011) we can hypothesise that the dominance of familiarity will apply to the face evaluation context where social judgements will be much more consistent and less variable for different images of familiar rather than unfamiliar identities.

Face averaging

An approach that has been proposed as the mechanism behind the process of familiarisation involves extracting summary statistics from faces or averaging images of the same identity together (Jenkins & Burton, 2011; Kramer, Ritchie, & Burton, 2015). This was originally introduced by Francis Galton (1879) who used photographic superimposing techniques and extended to create digitally blended composite faces by Benson and Perrett (1993). In order to construct an average image from a number of exemplar images, we morph the average face shape of those images represented by the mean xy-coordinates of a number of manually-aligned facial landmarks and the average face texture, represented by the mean RGB values at each pixel of the image (see Burton, Jenkins, Hancock, & White, 2005 for further details on average construction). The appeal of the averaging process comes from the fact that as more images are incorporated into the average, all superficial and temporary information is averaged-away while all critical and identity-diagnostic information is preserved making the average a stable identity representation. People have been shown to represent the average of face sets containing both familiar and unfamiliar faces (Kramer et al., 2015; Neumann, Schweinberger, & Burton, 2013) and their advantages have been shown in both human and computer identity recognition (Jenkins & Burton, 2008; White, Burton, Jenkins, & Kemp, 2014). Average-based computational systems, for example, have been shown to outperform systems trained on single exemplars only and in humans, the time to recognise someone as familiar can be decreased by including more images in the average (Burton et al., 2005).

While face averages have been shown to be a more accurate representation of identity and improve recognition performance, they possess some qualitative differences from exemplar images. Something that has been pointed out by both Galton (1879) and Benson and Perrett (1993) is that average images are much smoother in texture and have a rather soft-focus effect to them. Judging by the improvement in face recognition with averages, it can be safely assumed that successful recognition does not depend on the fine details in the texture and complexion of the face, however this blurring effect, combined with other averaging artefacts such as the removal of blemishes and temporary imperfections of the skin, could influence social evaluation in terms of attractiveness, trustworthiness and even distinctiveness. With averages being proposed as an alternative form of photo ID (White et al., 2014) and evidence of the great impact of first impressions on people's choices and behaviours (Todorov et al., 2015 for a review) it is important to explore the social information conveyed by average images. Some evidence for the possible effect of averages on social evaluation comes from studies using facial composite faces (Langlois & Roggman, 1990; Langlois et al., 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 are evaluated in a qualitatively different way than normal exemplar images.

Overview of studies

The experiments in this chapter aim to explore gender differences in the relationships between social evaluation traits based on Sutherland, et al. (2015). They also investigate the social information conveyed by average morphs of normal exemplar images. We further focus on the role of familiarity in social evaluation and compare the consistency between ratings of familiar and unfamiliar faces. What is different here from previous first impressions studies is that we make use of within-person variability by using different

images of the same identity and that these images are more natural and vary in emotional expression, lighting, pose, etc., therefore represent everyday social interactions more accurately. Experiment 1 compares ratings of attractiveness, trustworthiness, dominance, distinctiveness, and extraversion for average and exemplar images of unfamiliar identities and Experiment 2 demonstrates how these ratings change with familiarity. Finally, we report an additional analysis that aims to explore ratings variability in familiar and unfamiliar faces using data from the previous two experiments.

2.2. Experiment 1

Introduction

In the first experiment, we are interested in the relationship between social evaluation traits for unfamiliar identities. Evidence from the social stereotypes literature implies that faces are automatically assigned to a specific category (e.g. males vs females) and then category-related information is used to form a first impression of that identity (Secord, 1959). As most influential face evaluation models (Oosterhof & Todorov, 2008; Walker & Vetter, 2009) are more interested in establishing the physical information in the face people use to inform their judgements, they use images of male and female identities together which leaves any possible differences in ratings, due to gender, undetected. Nevertheless, key findings from Sutherland et al. (2015) show that female faces high in trustworthiness are perceived as significantly less dominant, whereas no such pattern is evident for ratings of male identities.

Experiment 1 further aims to explore the process of averaging which entails morphing a number of images of the same identity together in order to construct a more optimal representation of that particular identity. This is due to the fact that averaging preserves any identity-specific and diagnostic information while washing away the effects of any short-lived sources of face variability such as emotional expression, lighting, or temporary health changes. This aspect of averaging has been shown to be beneficial for face

recognition in both familiar and unfamiliar faces (Burton et al., 2005; White et al., 2014), however, it presents with further artefacts such as excessive blurring or smoothing of face texture which has been shown to affect the way these images are socially evaluated (Little & Hancock, 2002). Therefore, we are also interested in establishing the social information conveyed by average images and how this information compares with ratings of normal exemplar images of the same identity.

Method

Participants

A total of 27 participants (3 male, M = 21 years, age range: 18-31) from the University of Aberdeen took part in the study. All had normal or corrected-to-normal vision and received payment or course credit for their participation. Participants provided informed consent prior to their participation in accordance with the ethical standards stated in the 1964 Declaration of Helsinki. Experimental procedures were also approved by the ethics committee of the University of Aberdeen psychology department.

Materials

A total of 200 face images were used as experimental stimuli. These included four different exemplar images of 40 unfamiliar identities (20 male & 20 female) as well as an average of those four images for each identity (detail of average construction to follow). Exemplar images were downloaded from a Google Image Search by entering the name of the identity 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 the 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 Figure 2.1 for examples and Burton et al., 2005 for further details). Images were morphed using a custom MATLAB software (Kramer, Jenkins, & Burton, 2016).

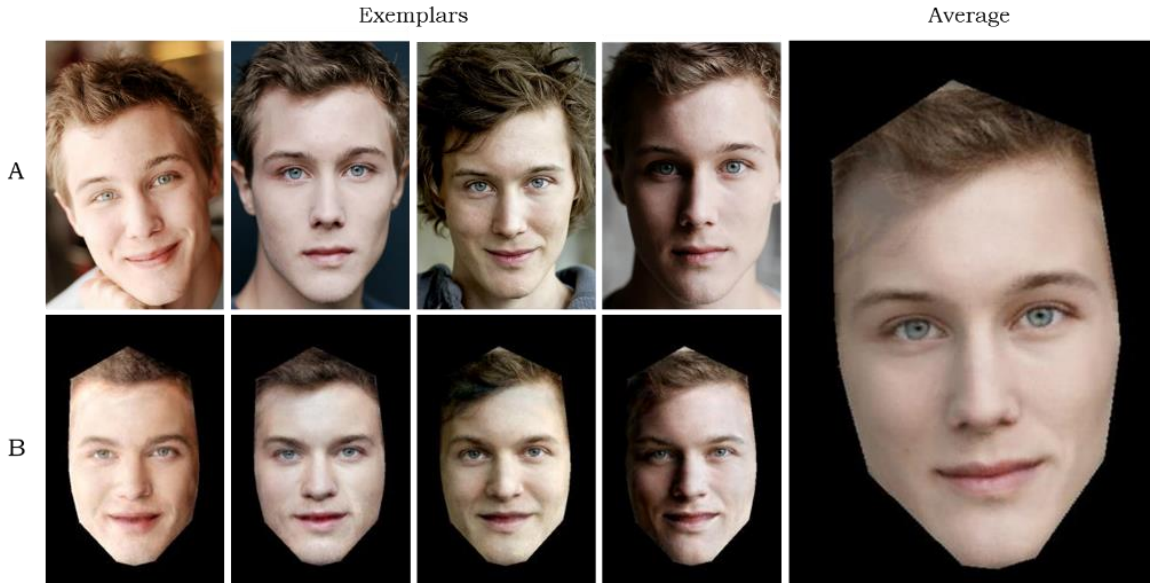

Figure 2.1. Four exemplar images of a single identity. (A) shows the original images and (B) shows the results of these image being morphed to a standard shape. The larger image on the right is the average image of these shape-standardized images.

Design and procedure

Participants were tested individually in a lab at the University of Aberdeen, equipped with a standard PC running MATLAB R2014a. Stimuli were displayed on an 18-inch monitor and the experimental program was written in MATLAB using functions from the Psychophysics Toolbox (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). Participants

were asked to rate all 200 images for attractiveness, trustworthiness, distinctiveness, extraversion, and dominance on a scale from 1 (not at all) to 9 (extremely). Each face was presented individually at the centre of the screen with the rating scales positioned below the image and participants rated that face for all attributes at the same time. Face presentation order was randomised.

Results and discussion

First impressions across genders

Tables 2.1 and 2.2 show Pearson's correlations between social traits for male and female identities respectively. Attractiveness and trustworthiness were positively correlated for both male (r (100) = .76, p < .001) and female (r (100) = .50, p < .001) faces, however this relationship was significantly stronger for male faces (z = 3.11, p < .01). Positive correlations were also found between attractiveness and distinctiveness (r (100) = .41, p < .001 and r (100) = .49, p < .001 for male and female faces respectively) as well as between attractiveness and extraversion (r (100) = .31, p < .01 and r (100) = .46, p < .001 for male and female faces respectively). Differences in the relationships between social attribute judgements concern ratings of dominance where there were significant negative correlations between attractiveness and dominance (r (100) = -.26, p < .01) as well as between trustworthiness and dominance (r (100) = -.43, p < .001) for female faces but there were no such correlations for male faces. This demonstrates that female faces perceived as more dominant are also perceived as less attractive and trustworthy, whereas the perception of dominance in male faces does not seem to affect the perception of their attractiveness and trustworthiness.

Results show clear gender differences in the correlations between social attributes. The correlation between trustworthiness and attractiveness was much higher for male faces than female faces. Such findings imply that there might be a cut-off point for female faces where highly attractive faces start to be perceived as untrustworthy. Moreover, there was a negative correlation between attractiveness and dominance as well as between trustworthiness

and dominance for female identities, demonstrating that more dominant-looking female faces are perceived as less attractive and trustworthy, however no such relationship was found for male faces. Such results support gender stereotype studies as well as findings from Sutherland et al. (2015).

Table 2.1. *Mean Social Attribute Ratings and Correlations Between Social Traits for Unfamiliar Male Identities.*

	M (SD)	1	2	3	4	5
1. Attractiveness	4.38 (.92)	-	.76***	.41***	.31**	.19
2. Trustworthiness	4.57 (.66)		-	.15	.42***	-.18
3. Distinctiveness	4.90 (.49)			-	.07	.21*
4. Extraversion	5.25 (.78)				-	.25*
5. Dominance	4.95 (.50)					-

Table 2.2. *Mean Social Attribute Ratings and Correlations Between Social Traits for Unfamiliar Female Identities.*

	M (SD)	1	2	3	4	5
1. Attractiveness	5.00 (.71)	-	.50***	.49***	.46***	-.26**
2. Trustworthiness	4.95 (.68)		-	.14	.42***	-.43***
3. Distinctiveness	5.07 (.46)			-	.22*	.41***
4. Extraversion	5.48 (.55)				-	.18
5. Dominance	5.17 (.53)					-

First impressions from averages

Mean ratings by condition for all social traits are shown in Figure 2.2. A 5 x 2 within subjects ANOVA with factors social trait (attractiveness, trustworthiness, distinctiveness, extraversion, and dominance) and image type (average image vs mean of exemplars) revealed a significant main effect

of trait (F (4, 156) = 9.67, p < .001, η_p^2 = .20), but not image type (F (1, 39) = 2.27, p > .05, η_p^2 = .06). These main effects were qualified by a significant interaction (F (4, 156) = 8.72, p < .001, η_p^2 = .18). Simple main effects showed a significant effect of image type for attractiveness (F (1, 195) = 18.61, p < .001, η_p^2 = .09) and trustworthiness (F (1, 195) = 8.77, p < .01, η_p^2 = .04) ratings, where average images were rated as significantly more attractive and more trustworthy than the exemplar images.

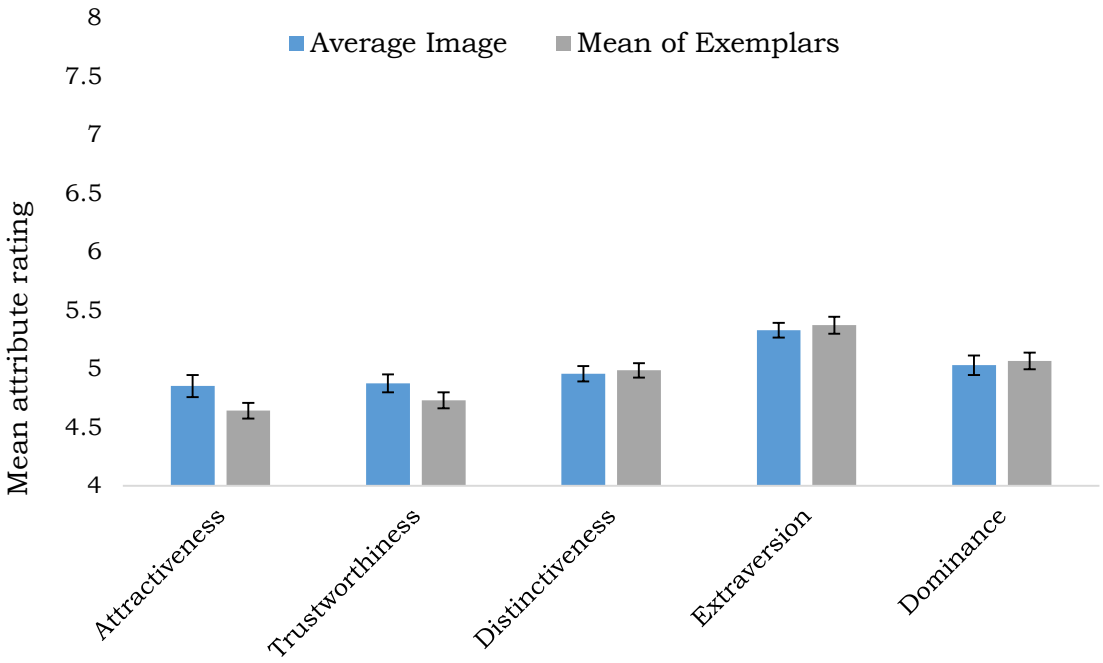

Figure 2.2. Mean ratings of exemplar and average images for unfamiliar identities across all social attributes. Error bars represent within-subjects standard error (Cousineau, 2005).

Analysing the data separately for male and female identities revealed that this effect was solely driven by the female identities. Figures 2.3 and 2.4 show the mean ratings of average and exemplar images for all social traits separately for male and female identities. The same 5 x 2 within subjects ANOVA for female identities showed significant main effects of trait (F (4, 76) = 3.35, p < .01, η_p^2 = .15) and image type (F (1, 19) = 19.64, p < .001, η_p^2 = .51) as well as significant interaction between the two (F (4, 76) = 12.46, p < .001, η_p^2 = .40). Simple main effects revealed the same pattern of results with

average images rated as more attractive (F (1, 95) = 54.42, p < .001, η_p^2 = .36) and trustworthy (F (1, 95) = 19.01, p < .001, η_p^2 = .17) than exemplar images. Analysis for male identities showed a main effect of trait only (F (4, 76) = 8.11, p < .001, η_p^2 = .30) but no main effect of image type (F (1, 19) = 2.79, p > .05, η_p^2 = .13) or interaction between the two (F (4, 76) = 1.26, p > .05, η_p^2 = .06).

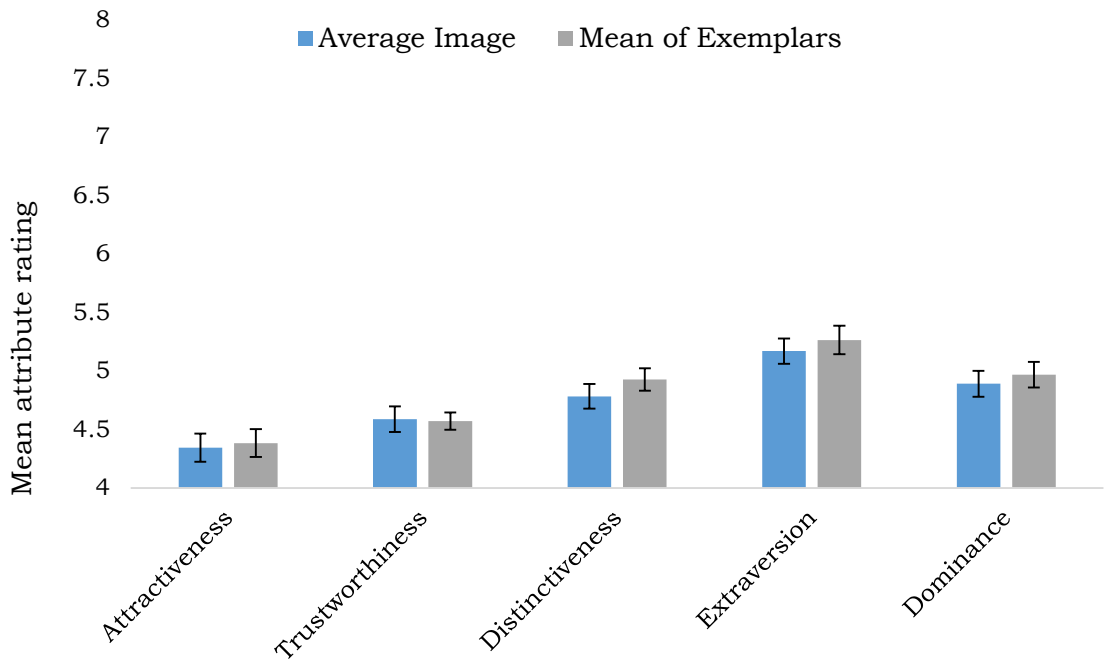

Figure 2.3. Mean ratings of exemplar and average images for unfamiliar male identities. Error bars represent within-subjects standard error (Cousineau, 2005).

Investigating the differences in social attribute judgements for average and exemplar unfamiliar faces showed that overall the physical average image corresponds to a social average in term of distinctiveness, extraversion, and dominance. Looking at these results separately for male and female faces showed clear gender differences where the physical average conveys the same social information as the exemplar faces about male, but not female faces. Averages of female faces were perceived as more attractive and trustworthy than single exemplars, possibly because the process of averaging can even out the skin tone and blur out any imperfections in the skin. Attractiveness

and trustworthiness are often found to be highly correlated in social judgements supporting the idea of a 'halo effect' (Dion et al., 1972), therefore changes in attractiveness can explain differences in trustworthiness ratings as well. Next, we introduce familiarity and investigate the effect of averaging as well as gender differences in social evaluation for familiar faces. This will allow us to establish any changes in social perception and compare trait attribution for familiar and unfamiliar faces.

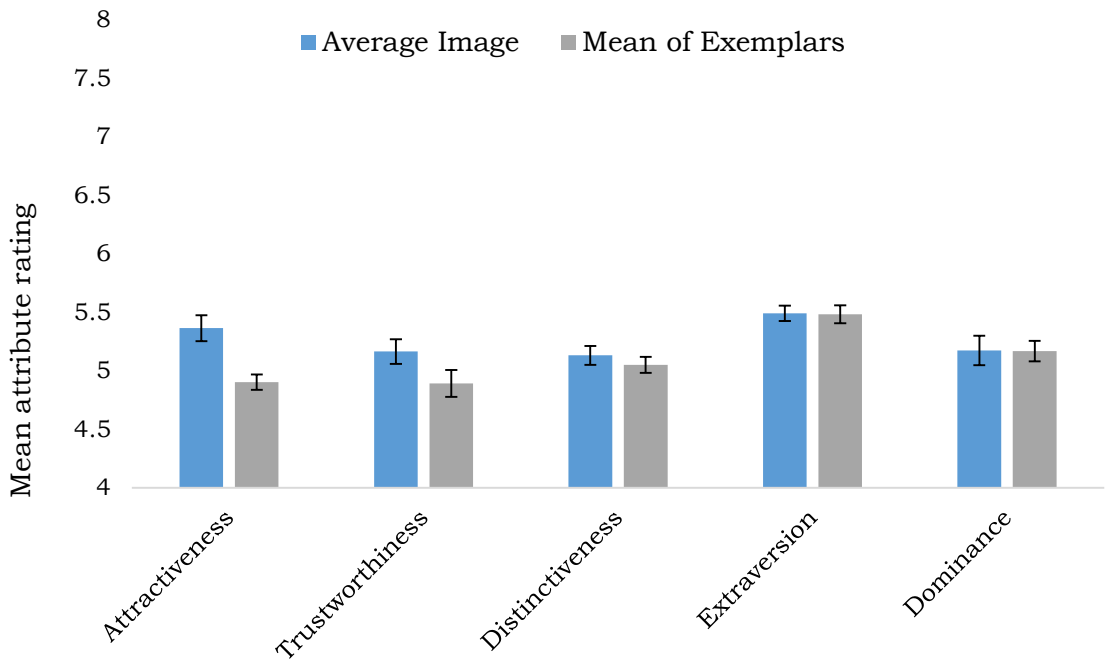

Figure 2.4. Mean ratings of exemplar and average images for unfamiliar female identities. Error bars represent within-subjects standard error (Cousineau, 2005).

2.3. Experiment 2

Introduction

Our next experiment explores how the relationship between social traits and the effects of averaging change with familiarity. Evidence from studies manipulating familiarity show that faces that have been seen before as well as faces resembling familiar identities closely are perceived more

favourably than unfamiliar faces (Rhodes et al., 2001; Zebrowitz, 1996). Lewicki (1985), for example, demonstrated that people showed a more favourable attitude towards others whose faces resembled someone who had just treated them kindly and a rather negative attitude towards people whose faces resembled someone who had just irritated them. With face evaluation studies consistently reporting strong positive relationship between ratings of attractiveness and trustworthiness, it would be expected that familiar faces might be perceived as more attractive and trustworthy than unfamiliar faces.

Moreover, taking the idea of stereotypes in first impressions and Secord's social categorisation theory (1959) a step further, another prediction would be that while unfamiliar faces will be evaluated based on information about the social category they have been assigned to, familiarity and actual knowledge about the person would overshadow this mechanism for familiar faces. As we become familiar with a particular identity, we are exposed to the different ways this person may look as well as to their personality. Therefore, not only are we forming a mental representation of that person's appearance, which is similar to the process of averaging, but we are also forming a stable social impression of that identity's personality. Based on this idea, we would expect different images of the same familiar identity to be very similarly evaluated as ratings would be based on prior knowledge and experience rather than the physical properties of the images.

Method

Participants

A total of 27 participants (5 male, M = 20.8 years, age range: 18-27) from the University of York took part in the study. All had normal or corrected-to-normal vision and received payment or course credit for their participation. Participants provided informed consent prior to their participation in accordance with the ethical standards stated in the 1964 Declaration of Helsinki. Experimental procedures were also approved by the ethics committee of the University of York psychology department.

Materials

A total of 200 face images were used as experimental stimuli. These included four different exemplar images of 40 familiar identities (20 male & 20 female) as well as an average of those four images for each identity (constructed the same way as the ones in Experiment 1, see Figure 2.5 for examples). Exemplar images were downloaded from a Google Image Search by entering the name of the identity 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. These were all 'ambient' images of world-known celebrities, which 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.

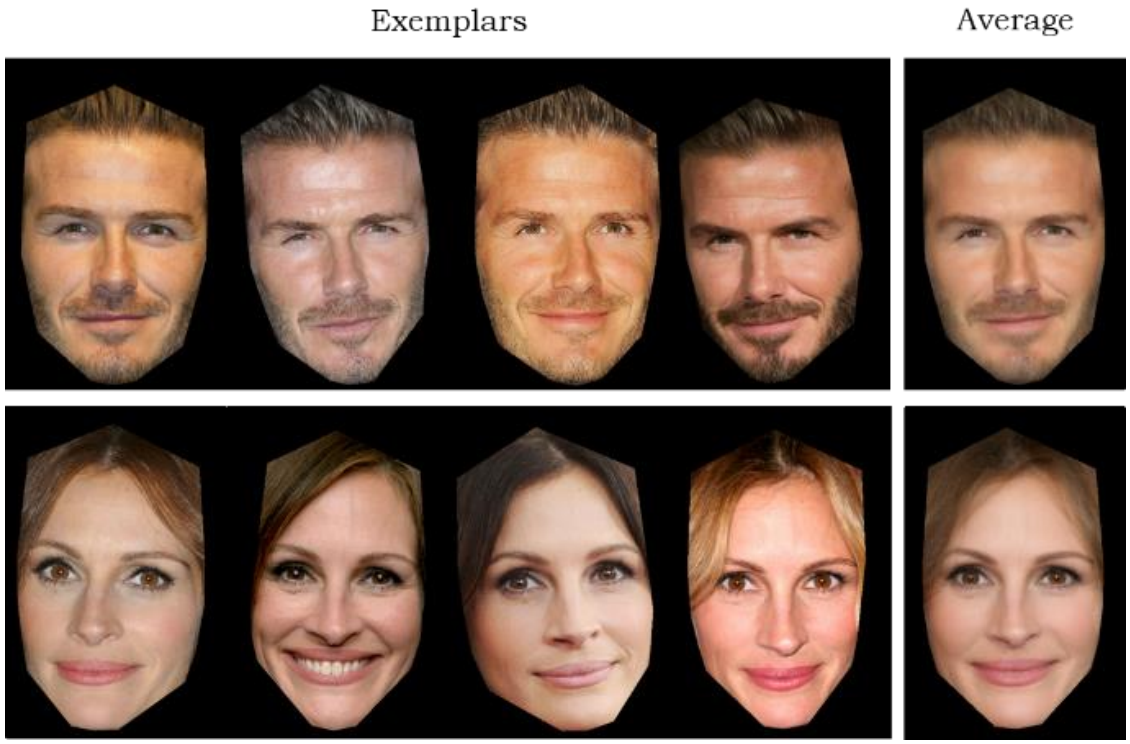

Figure 2.5. Exemplar and average image examples for familiar identities. Images on each row are of the same identity.

Design and procedure

The design and procedure of the this experiment were exactly the same as those of Experiment 1 with the only difference being the type of face

images used. While participants rated identities unfamiliar to them in Experiment 1, here participants were asked to rate familiar faces.

Results and discussion

First impressions across genders

Tables 2.3 and 2.4 show Pearson's correlations between social attributes for male and female faces respectively. Attractiveness and trustworthiness were positively correlated for both male (r (100) = .56, p < .001) and female (r (100) = .76, p < .001) faces, however this relationship was significantly stronger for female identities (z = 2.53, p < .05). Positive correlations were also found between attractiveness and distinctiveness (r (100) = .43, p < .001 and r (100) = .68, p < .001 for male and female faces respectively) as well as between attractiveness and dominance (r (100) = .53, p < .001 and r (100) = .32, p < .01 for male and female faces respectively). Trustworthiness was positively correlated with distinctiveness (r (100) = .40, p < .001 and r (100) = .60, p < .001 for male and female faces respectively) and extraversion r (100) = .33, p < .01 and r (100) = .21, p < .05 for male and female faces respectively) for both male and female faces.

Table 2.3. *Mean Social Attribute Ratings and Correlations Between Social Traits for Familiar Male Identities.*

	M (SD)	1	2	3	4	5
1. Attractiveness	4.38 (.92)	-	.56***	.43***	.38***	.53***
2. Trustworthiness	4.57 (.66)		-	.40***	.33**	-.06
3. Distinctiveness	4.90 (.49)			-	.37***	.45***
4. Extraversion	5.25 (.78)				-	.09
5. Dominance	4.95 (.50)					-

Table 2.4. *Mean Social Attribute Ratings and Correlations Between Social Traits for Familiar Female Identities.*

	M (SD)	1	2	3	4	5
1. Attractiveness	5.00 (.71)	-	.76***	.68***	.15	.32**
2. Trustworthiness	4.95 (.68)		-	.60***	.21*	-.10
3. Distinctiveness	5.07 (.46)			-	.38***	.49***
4. Extraversion	5.48 (.55)				-	.01
5. Dominance	5.17 (.53)					-

Results showed a positive correlation between attractiveness and dominance for both male and female faces and no significant correlations between trustworthiness and dominance. Comparing these results with those for unfamiliar faces it is clear that familiarity overwrites the negative relationship between dominance and both attractiveness and trustworthiness. Once we become familiar with a dominant female face it is no longer perceived as less attractive and trustworthy. On the contrary, familiar faces perceived as more dominant are also perceived as more attractive. Overall, comparing the correlations between attributes for familiar and unfamiliar faces we can see that familiarity acts as an equaliser as it reduced the magnitude of gender differences leading to a similar pattern of correlations for both male and female faces.

First impressions from averages

Mean ratings by condition for all social traits are shown in Figure 2.6. A 5 x 2 within subjects ANOVA with factors social trait (attractiveness, trustworthiness, distinctiveness, extraversion, and dominance) and image type (average image vs mean of exemplars) showed no significant main effect of trait (F (4, 156) = 2.22, p > .05, η_p^2 = .05) or image type (F (1, 39) < 1, p > .05, η_p^2 = .01). These main effect were qualified by a significant interaction (F (4, 156) = 22.10, p < .001, η_p^2 = .36). Simple main effects showed a significant effect of image type for attractiveness (F (1, 195) = 18.32, p < .001, η_p^2 = .09) and trustworthiness (F (1, 195) = 21.88, p < .001, η_p^2 = .10) ratings where

average images were rated as significantly more attractive and more trustworthy than the exemplar images. Simple main effects also showed that average images were rated as less distinctive (F (1, 195) = 11.61, p < .001, η_p^2 = .06) and dominant (F (1, 195) = 9.08, p < .01, η_p^2 = .04) than exemplar images.

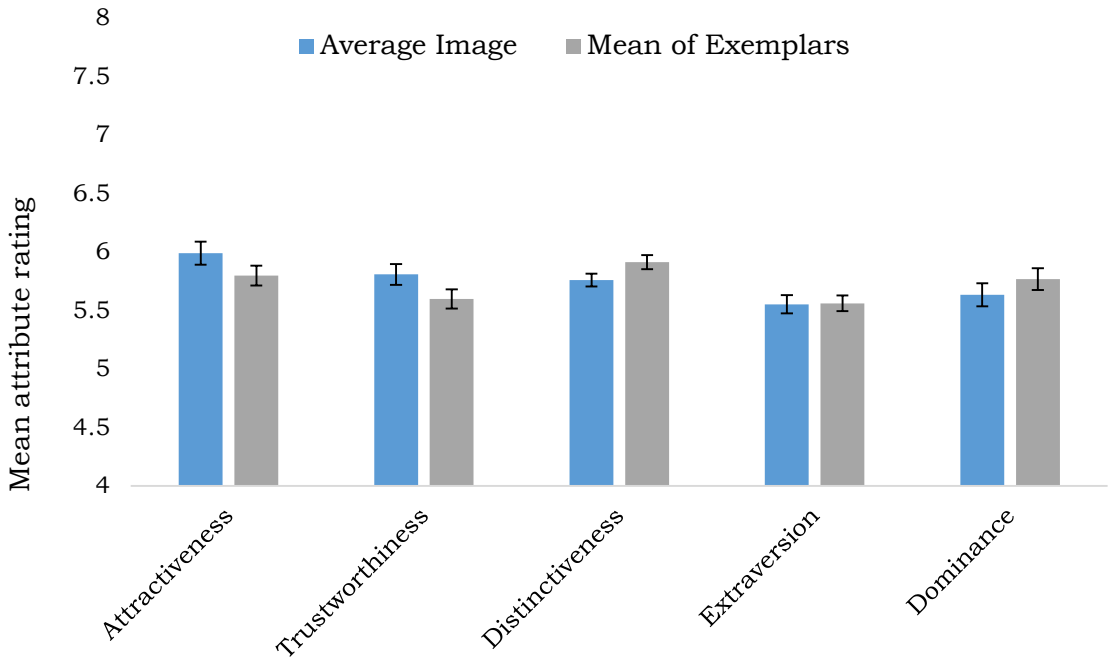

Figure 2.6. Mean ratings of exemplar and average images for familiar identities across all social attributes. Error bars represent within-subjects standard error (Cousineau, 2005).

Figures 2.7 and 2.8 show the mean ratings of average and exemplar images for all social traits separately for male and female identities. The same 5 x 2 within subjects ANOVA for female identities showed significant main effects of trait (F (4, 76) = 11.65, p < .001, η_p^2 = .38) but not of image type (F (1, 19) = 3.40, p > .05, η_p^2 = .15). The interaction between trait and image type was found significant (F (4, 76) = 26.43, p < .001, η_p^2 = .58) and simple main effects revealed that average images were rated as more attractive (F (1, 95) = 37.98, p < .001, η_p^2 = .29) and trustworthy (F (1, 95) = 23.62, p < .001, η_p^2 = .20) than exemplar images. Same analysis for male identities showed a main effect of trait (F (4, 76) = 4.74, p < .01, η_p^2 = .20) but no main effect of image type (F (1, 19) = 1.04, p > .05, η_p^2 = .05). The interaction between trait

and image type was found significant (F (4, 76) = 6.77, p < .001, η_p^2 = .26) with simple main effects showing that average images were rated as less distinctive (F (1, 95) = 13.78, p < .001, η_p^2 = .13) and less dominant (F (1, 95) = 4.43, p < .01, η_p^2 = .04) than exemplar images.

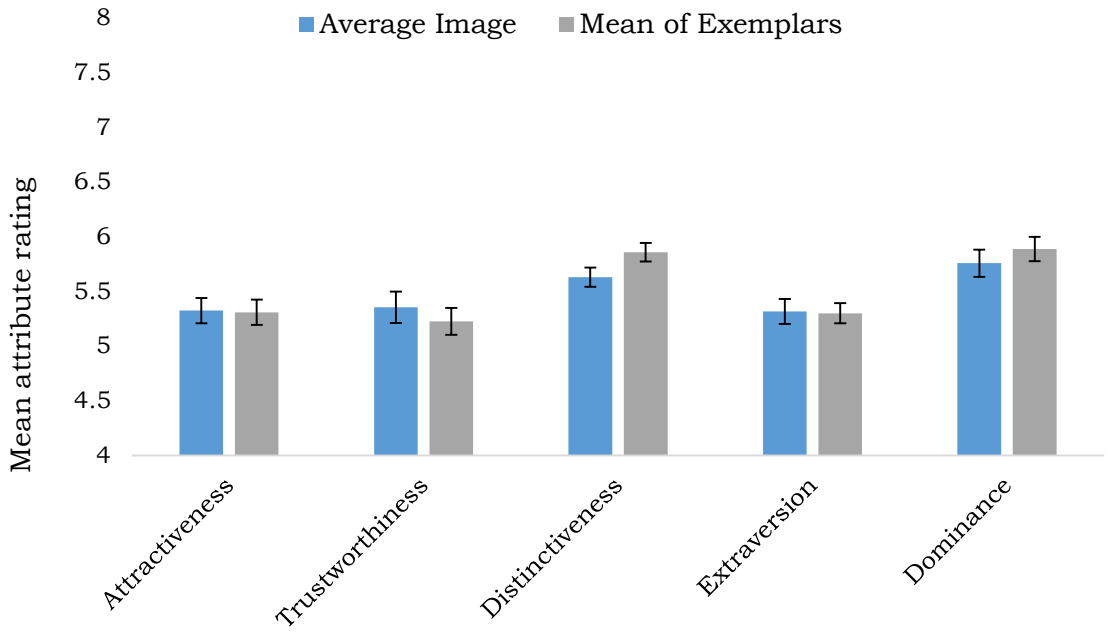

Figure 2.7. Mean ratings of exemplar and average images for familiar male identities. Error bars represent within-subjects standard error (Cousineau, 2005).

Results showed that the physical average of familiar faces conveys different social information compared to exemplar images of the same identities. Again, there were clear gender differences where averages of female faces are perceived as more attractive and trustworthy than exemplar images and averages of male faces are perceived as less distinctive and dominant than exemplar images. All these differences can be a result of the smoothing and blurring involved in creating an average image. Gender differences in the relationships between social traits imply that particular social traits might be of higher importance or relevance for male and female identities. Attractiveness and trustworthiness might be more diagnostic for female identities, whereas dominance and distinctiveness might be more diagnostic

for male identities. Such an interpretation is more compatible with general social evaluation models rather face evaluation models, which is not surprising given that all face evaluation models are based on judgements of unfamiliar faces.

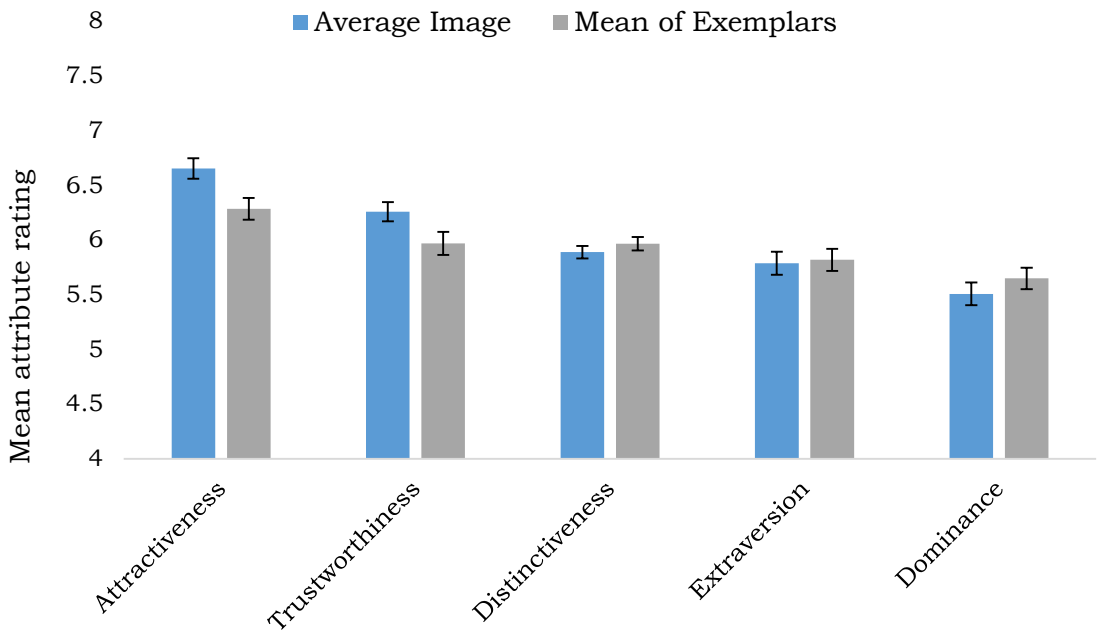

Figure 2.8. Mean ratings of exemplar and average images for familiar female identities. Error bars represent within-subjects standard error (Cousineau, 2005).

2.4. Comparing Ratings of Familiar and Unfamiliar Identities in Social Face Space

Finally, we aim to compare social evaluations of familiar and unfamiliar identities in a more direct and comprehensive way. Integrating the rating data from the first two experiments together we created a 5-dimensional social attribute space where the position of each image depended on its ratings of attractiveness, trustworthiness, dominance, distinctiveness, extraversion, and dominance. This way we can explore how close in this social space different images of the same person are located. The prediction is that exemplar images of familiar identities will be located much closer together, as social attribute ratings will not depend on the physical properties of the

image. Once we are familiar with a person, differences in appearance are less likely to have an effect on social judgements attribution as the effect of familiarity takes over.

To test this idea we take sets of social attribute ratings for two randomly chosen exemplars per identity and compare their correspondence using Procrustes analysis (Gower, 1975). This method tries to transform 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. This is achieved through translation, reflection, rigid rotation, and scaling of the coordinate matrices which preserves the location of the images relative to one another. 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 one, the location values within each trait were shuffled (Jackson, 1995) and for the second one, 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. For each method of chance, the proportion of iterations where the chance goodness-of-fit was smaller than or equal to the observed value provided the significance level of the test. It is also important to note that a lower sum of squares distance (i.e., a lower value for goodness-of-fit) means a better fit. Table 2.5 shows the mean fit and fit 'by chance' for both familiar and unfamiliar identities.

Table 2.5. *Mean Fit of Data as well as Fit from the Chance Measures for Familiar and Unfamiliar Identities*

	Familiar Faces	Unfamiliar Faces
Mean fit of data (SD)	.39 (.05)	.62 (.05)
Mean fit for Shuffle1 (SD)	.93 (.03)	.93 (.03)
Mean fit for Shuffle2 (SD)	.94 (03)	.93 (.03)

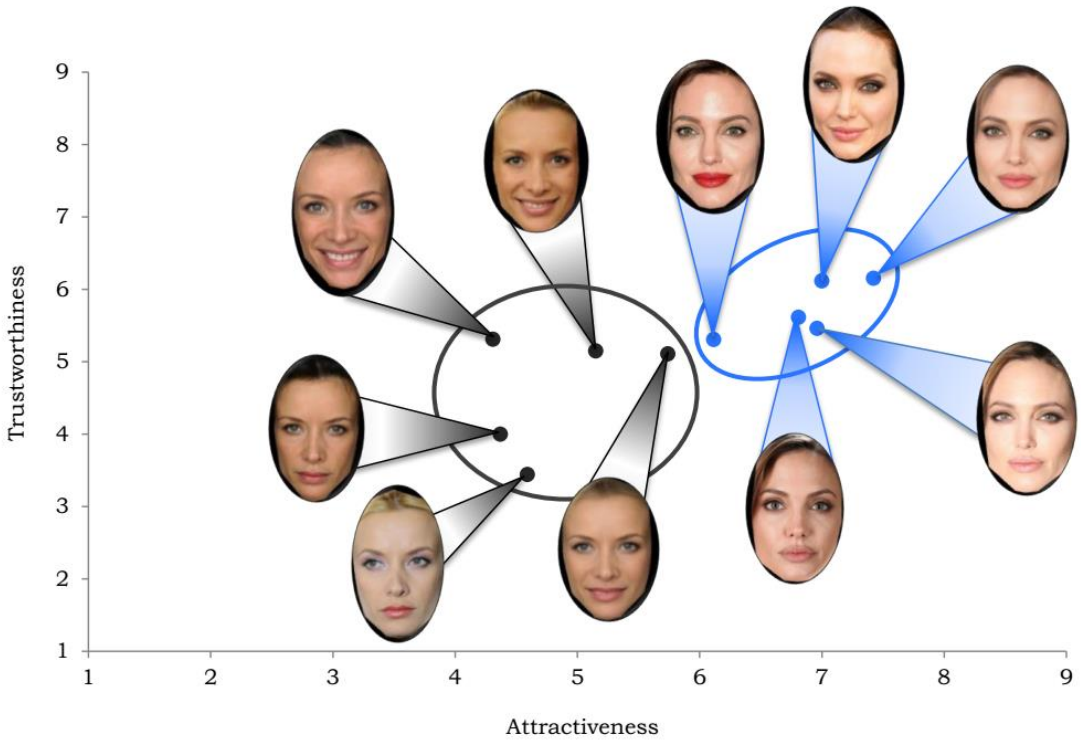

Figure 2.9. Example of the location of images in two-dimensional face space. Points' coordinates reflect real data for the familiar and unfamiliar faces sets.

The analysis shows a much better fit for both types of faces compared to chance and more importantly, a much better fit for familiar rather than unfamiliar faces. This implies that images of the same familiar identity are located much closer to one another in this social face space. A simplified example is presented in Figure 2.9 where each point represents a different

image and different colours represent different identities (black – unfamiliar identity, blue – familiar identity). This is only a simplified two-dimensional space example using ratings of attractiveness and trustworthiness. The figure illustrates that images of the familiar identity lie much closer together than the images of the unfamiliar identity. This therefore implies that familiarity takes over personality judgements as social ratings 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.

2.5. General Discussion

The experiments in this chapter aimed to investigate gender differences in face evaluation and explore the social information conveyed by a physical average image created through morphing different images of the same identity together. We were also interested in the possible effects of familiarity in social evaluation and whether they will outweigh the effects due to the physical properties of the images. Results showed clear gender differences in the relationship between attractiveness and dominance as well as trustworthiness and dominance with images high in dominance perceived as less attractive and less trustworthy for female identities, while no such relationships were found for male identities. This specific pattern of results was also affected by familiarity as we found a strong positive relationship between dominance and attractiveness for both male and female familiar identities. Finally, we show that the process of averaging and its associated artefacts such as blurring and smoothing of texture can have a significant effect on ratings of both familiar and unfamiliar faces. Taking ratings of all attributes together and projecting familiar and unfamiliar faces in a social face space demonstrated that ratings of different images of the same familiar identity are much less varied than ratings of unfamiliar identities.

Findings of the high attractiveness and trustworthiness ratings attributed to average images are somewhat consistent with those of Little and Hancock (2002) who also showed that composite images were perceived as

more attractive than exemplar images. However, they used male identities and we see these differences for female identities only. This could be due to the number of exemplar images that go into the average, as well as the fact that we were morphing different images of the same identity together. In their original study Little and Hancock used composites made up from 3, 6 and 12 faces and found significantly higher attractiveness ratings even for the 3-exemplar average. As we were morphing different images of the same person it is possible that the blurring effect is not as strong as when morphing images of different people together so we might observe a similar pattern of results for male identities if more images were included in the average. While the smoothing and blurring effect is rather small it seems like it is sufficient to bring about changes in social perception for female identities, implying that the face change threshold for female images might be lower than the one for male identities. Furthermore, introducing familiarity replicates the findings for female identities and accentuates the effect of the averaging process even further for male identities as we find that male averages are perceived as less distinctive and less dominant than the exemplar images. While it is clear how the blurring and smoothing can cause all of these changes in perception, the strong differentiation of important traits for male and female faces is particularly interesting. It seems like changes in female faces are highly diagnostic of attractiveness and trustworthiness judgements, whereas ratings of dominance and distinctiveness are more susceptible to changes in the male face. Such findings fit better with general social evaluation models which propose that the valence dimension reflects femininity and the dominance dimension reflects masculinity (Cuddy et al., 2008).

Exploring the correlations between social attributes for unfamiliar identities showed a much stronger relationship between attractiveness and trustworthiness for male than female identities. In general, the association between attractiveness and trustworthiness is commonly reported in social evaluation studies and reflects the 'halo effect', where positive qualities are attributed to attractive faces (Dion et al., 1972). A gender difference here is surprising, although it could imply some halo effect cut-off point after which

very attractive female faces are perceived as less trustworthy. Evidence for this interpretation comes from a study by Sofer et al. (2015) who used an average face made up of 92 images and an attractive average made up of the 15 most attractive images and created a continuum of images in-between using morphing. Their findings show that while attractiveness and trustworthiness increase together at the lower end of the spectrum, their relationship seems to be inverted at the higher end with the most attractive faces rated as less trustworthy (see Figure 1.6 on page 32).

Another surprising 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 (Langlois & Roggman, 1990; see Rhodes, 2006 for a review; Said & Todorov, 2011). Contrary to our findings, these studies report that a more typical face is evaluated more favourably and perceived as more attractive. Evolutionary theories explain this with attractiveness being an index of good health and 'good genes' as well as through the concept of averageness which describes how closely a certain face resembles the faces we encounter in everyday life. According to Thornhill and Gangestad (1993, 1999; Scheib, Gangestad, & Thornhill, 1999) a face close to the average is a signal for 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 the 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 experiments. 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.

Other key correlations that show clear gender differences concern the traits related to valence (attractiveness and trustworthiness) and dominance. Our results show strong negative relationships between these traits for female faces only. Such a pattern of results fits well with social stereotype studies and Sutherland et al. (2015) in particular 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 could reflect a 'backlash' effect which has been used to explain findings that more assertive and dominant women are received more negatively and are less likely to be successful at job interviews (Heilman, 2001; Rudman & Glick, 2001, although see Sczesny, Spreemann, & Stahlberg, 2006 for a different pattern of results when applying for a masculine-typed occupation). Another possibility is that dominance is 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.

Familiarity also brought about some significant changes in social evaluation as it seemed to equate and counteract most of the gender differences found for unfamiliar identities. Not only did familiarity present a similar pattern of results for male and female identities, but it also completely reversed the negative relationship between attractiveness and dominance for female identities. This, together with its effect on the way average images are perceived, indicates a possible change in the interpretation of these fundamental social traits, and especially dominance. It seems like dominance has a very negative connotation when it comes to unfamiliar identities (females in particular), whereas dominance is a favourable and desirable trait in familiar identities. Furthermore, the influence of familiarity is even clearer in our final analysis which demonstrates that different images of the same familiar identity are located much closer together in a face space of social traits. This implies that familiarity takes over differences in the physical properties of images as people's ratings are rather based on prior knowledge

and experience. It should be noted, however, that completely unfamiliar and familiar faces are just the two ends of the spectrum so utilising multiple levels of familiarity will improve our understanding of its effects to a greater extend.

The experiments in this chapter showed clear gender differences in the relationships between the fundamental social evaluation dimensions and that these differences are somewhat diminished after we introduce familiarity. Such findings extend existing first impressions literature by providing a more comprehensive comparison between the process of face evaluation for male and female identities. This contributes to our understanding of first impressions formation as most face evaluation models use primarily male faces (Oosterhof & Todorov, 2008) or extract information from both male and female faces together (Walker & Vetter, 2009) without considering that they might be evaluated differently. We also incorporate the effects of familiarity which might be seemingly counter-intuitive when it comes to zero-acquaintance first impressions. Nevertheless, we demonstrate how familiarity acts as an equaliser of gender differences in social evaluation and that it outweighs within-person variability not only in the context of identity recognition, as shown by previous research (Jenkins et al., 2001), but also in the context of social face evaluation.

Chapter 3 – Within-Person Variability in Social Evaluation

3.1 Introduction

The human face is an extremely rich stimulus and its perception is critical in the social world. It can provide us with a wealth of information about age, gender, race, emotional state, and identity (Bruce & Young, 1986). Relying solely on facial information, people readily form stable first impressions within a few milliseconds (Todorov et al., 2009, 2010; Willis & Todorov, 2006) demonstrating that such processes are automatic and outside of conscious control. Social attribution is characterised by a high level of agreement between observers (Zebrowitz & Montepare, 2008; Zebrowitz-McArthur & Berry, 1987) implying that there is some physical information in the face they use to inform their judgements. This high agreement in attribute ratings does not necessarily mean that these attributions are accurate (Todorov et al., 2015). Nevertheless, understanding first impressions is important as they have been repeatedly shown to influence social outcomes in a variety of contexts including dating preferences (Little, Burt, & Perrett, 2006), voting choices (Ballew & Todorov, 2007; Olivola & Todorov, 2010a), eyewitness testimony (Mueller et al., 1984, 1988) and police line-up selection (Flowe & Humphries, 2011). Moreover, in a recent study Wilson and Rule (2016) found that viewers accurately predicted sentences of convicted criminals (including 'life' and 'death' sentences) based on the trustworthiness perceived from their images alone.

Existing face evaluation models

Todorov and colleagues have developed a data-driven computational model of first impressions which aimed to extract the physical information in the face that is used to form these social judgements (Oosterhof & Todorov, 2008, 2011; Todorov, Said, Engell, & Oosterhof, 2008). To inform their model the authors asked participants to rate face images for a variety of social attributes and then subjected these ratings to principal components analysis (PCA). A two-component solution showed that there were two fundamental

dimensions of social face evaluation, which the authors interpreted as trustworthiness/valence and dominance. Such results are consistent with other dimensional models including accounts of concept evaluation (Osgood et al., 1957) and of interpersonal perception (Wiggins, 1979) which have also been shown to rely on two orthogonal dimensions – affiliation and dominance. Mapping out the extracted information on computer-generated faces, Oosterhof and Todorov (2008) were able to create a continuum of faces that spanned the trustworthiness and dominance dimensions. These artificially created faces were then rated by another sample of participants in order to validate the manipulation.

Another model that aimed to explore the underlying facial characteristics people use to inform their social attribute judgements is the Basel Face Model (Walker & Vetter, 2009). It uses PCA to extract the physical variability in face images (see Blanz & Vetter, 1999 for further methodological details) and then a regression analysis to identify the perceptual dimensions corresponding to personality traits in multidimensional space. These dimensions are then used to manipulate original or novel faces in order to change the way they are perceived. The validity of this procedure was tested by creating pairs of face images in which the same base face was manipulated along a single dimension and then used in a 2AFC task (i.e. contrasting high and low values of a single dimension). The results demonstrated that people were able to identify the direction in which all attributes were manipulated at above chance levels, implying that the underlying characteristics responsible for differences in social attribute judgements had been successfully identified and manipulated.

Within-person variability and natural variation

Most social evaluation studies are based on tightly controlled images of different identities. Using a single image to represent each identity corresponds to a view that *individuals* are perceived as more or less trustworthy, dominant, and so forth. This position carries an implicit assumption that different images of any particular person will give rise to similar social attribute ratings. However, recent studies investigating within-

person variability, or differences in images of the same identity, present some challenging findings.

Evidence from the face recognition literature comes from Jenkins et al. (2011) who introduced the concept of 'ambient images'. They argued that images used for face perception studies should reflect the range of photos that are encountered naturally, for example in social and broadcast media. These images vary in a number of ways, including emotional expression, age, facial hair, and make-up as well as physical image differences such as camera angle and lighting direction. Use of such images has revealed some key findings which are not available when studying highly constrained pictures in which variability is regarded as noise, and controlled away (Burton, 2013). Utilising a card sorting task Jenkins et al. (2011) asked participants to sort 40 face photos into piles, one per person. When unfamiliar with the people in these photos, viewers tended to over-estimate the number of identities – reporting on average nine piles, when in fact only two identities were present. However, they made very few confusion errors – tending not to sort different people into the same pile. In this case, the difficulty for viewers was not telling people apart, but 'telling people together'. This effect is entirely due to within-person variability – unfamiliar viewers find it difficult to cohere together different images of the same person. Despite representing an important component of face processing, within-person variability is little studied.

As judging faces on socially important attributes is considered an automatic process and involves analysis of unfamiliar faces, it is possible that it is also affected by within-person variability. Jenkins et al. (2011) tested this idea by collecting attractiveness ratings for 20 different images of 20 identities and demonstrated that attractiveness evaluation is not stable, but varies very widely across different images of the same person. Further evidence comes from Todorov and Porter (2014) who collected ratings for five images of 20 different identities on a number of social attributes including attractiveness, trustworthiness, and extraversion. They showed that the variance in ratings

between images of the same person is comparable to (and sometimes exceeds) the variance of images between different people.

Findings from both studies imply that social evaluation does not depend solely on identity but also on the specific image used to represent any particular identity. This presents us with a novel possibility – just as it is possible to use between-person differences to manipulate people's perceptions, it should also be possible to manipulate perceptions of an individual's face, without changing their identity.

Overview of studies

This chapter aims to extend findings from previous face evaluation models by incorporating between- and within-person variability together (Experiment 3) as well as exploring the potential of purely idiosyncratic information to change social perceptions (Experiment 4). Our approach is novel in the use of great many natural and highly variable images of a few identities. We use a statistical analysis of both the shape and texture information in face images to capture within-person variability. We can then establish the image dimensions which contribute to the perception of different social attributes. This will allow us to 1) explore the underlying physical characteristics people use to inform their social judgements specifically for each identity, 2) manipulate the way these identities are perceived (Experiment 5), and 3) apply the same technique to novel images of the same identity (Experiment 6) in order to investigate whether the variability extracted from the original images can also be generalised to further images of the same identity. Furthermore, we explore some of the low-level image properties and face metrics and their predictive value for social evaluation (Experiment 7).

3.2 Experiment 3

Introduction

There is now evidence that different images of the same person can produce considerably different ratings on social dimensions (Jenkins et al.,

2011; Todorov & Porter, 2014). However, current face evaluation models are based on between-person differences, and do not take within-person variability into consideration. This raises the question of whether social attribution is based on identity only, or whether it also depends on the properties of specific images. In order to address this question we firstly integrate within- and between-person variability together which will allow us to compare the variance in these attributions between people (e.g. are some people consistently rated more or less trustworthy than others?) and within people (e.g. to what extent are pictures of the same person rated to be similarly trustworthy?). We will also explore whether this approach can provide sufficient variability to capture the underlying physical information in the face which is diagnostic for each social attribute.

Rather than using a single image to represent each identity, here we use 20 different images of 20 unfamiliar identities (referred to as the 20-20 set) and collected ratings of attractiveness, trustworthiness, and dominance for all 400 images. This allows us to examine the magnitude and spread of social attribute ratings within as well as between identities. We then performed statistical analysis on the face images, using principal components analysis, in order to establish the extent to which patterns of within- and between-person variability can be captured by physical variance in the images themselves. This provides us with a model linking image characteristics to people's ratings. In order to establish the validity of this novel approach, we then used the model to manipulate the way images were socially perceived. To anticipate the results we find that ratings of attractiveness are predicted to a large extent by between-person variability, i.e. some people are consistently rated more attractive than others. However, this is not the case for ratings of trustworthiness and dominance, implying that these social judgements rely as much on properties of the particular image as on properties of the person depicted.

Method

Participants

Images were rated by 20 participants (2 male, mean age = 20.1, age range = 18-24), all from the University of York. All had normal or corrected-to-normal vision and received payment or course credit for their participation. Participants provided informed consent prior to their participation in accordance with the ethical standards stated in the 1964 Declaration of Helsinki. Experimental procedures were also approved by the ethics committee of the University of York psychology department.

Stimuli

The image set consisted of 20 images of each of 20 unfamiliar people (10 men), 400 in all. These were foreign celebrities and associates, which ensures the availability of many images for each identity. All were unfamiliar to British viewers. Images were downloaded from an internet search by entering the name of the person and choosing images that were in full colour, broadly frontal, and with no parts of the face obscured by clothing or glasses. These were all naturally occurring or 'ambient' images (Jenkins et al., 2011) and captured a great amount of face variability due to lighting, pose, and expression for each identity (see Figure 3.1 for examples).

Design and procedure

The rating task was computer-based, and stimuli were displayed on an 18-inch LCD monitor. The experimental program was written in MATLAB and used functions of the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). Face images were presented individually at the centre of the screen with a rating scale positioned underneath. Participants were asked rate each image on a scale from 1 (not at all attractive/ trustworthy/ dominant) to 9 (extremely attractive/ trustworthy/ dominant) using a mouse-click. The task was self-paced, and inter-stimulus interval was 1s. Participants were not given detailed instructions as to how to interpret the ratings dimensions (attractiveness/ trustworthiness/ dominance) but were encouraged to rely on their "gut instinct" (cf Todorov, Mandisodza, Goren, &

Hall, 2005). Participants provided ratings for all 400 images. Each face was rated for all social attributes in sequence, and the order of the three ratings was randomised for each image to avoid carryover effects (Rhodes, 2006). Order of image presentation was also randomised individually for each participant.

Figure 3.1. Example ambient images of the same identity

Results and discussion

Inter-rater reliability was very high for all three social attributes (Cronbach's alphas above .90, Nunnally, 1978). Figure 3.2 shows mean ratings for all images on the three social judgements – displayed separately for male and female identities, and ranked by overall mean, separately for each judgement. The figure illustrates very interesting differences among social ratings. Consistent with much previous work (Jenkins et al, 2011; Todorov & Porter, 2014) there are very large differences in the ratings given to different images of the same person, and this is true for all the three judgements. For attractiveness, there seems to be a fairly clear difference between people – some people (and particularly men) are consistently rated as being less attractive than others, even within the context that there is some variability among different photos of the same individuals. However, this inter-person difference is much less evident for judgements of trustworthiness and dominance. In these cases, it is relatively hard to see much consistent effect of target person. Even when ordered by mean rating per target person, there is rather little pattern to see – the variability between people is relatively smaller than the variability within each identity.

These observations were confirmed by statistical analysis. Variances are shown in Table 3.1 (following Todorov & Porter, 2014). We compared within-person variance to between-person variance for each of the attributions, separately for men and women (F-test, two-tailed, $p = 0.05$ adjusted by Bonferroni correction for multiple comparisons). For male faces, there was significantly more variance in attractiveness ratings between individuals than within individuals ($F(9, 90) = 10.37$, $p < .05$). For all other comparisons, there was no significant difference in the variance accounted for by between and within-individual scores. In particular, there is no hint to any between-person advantage for trustworthiness and dominance – where within-person variance was numerically larger than between person variance in every case.

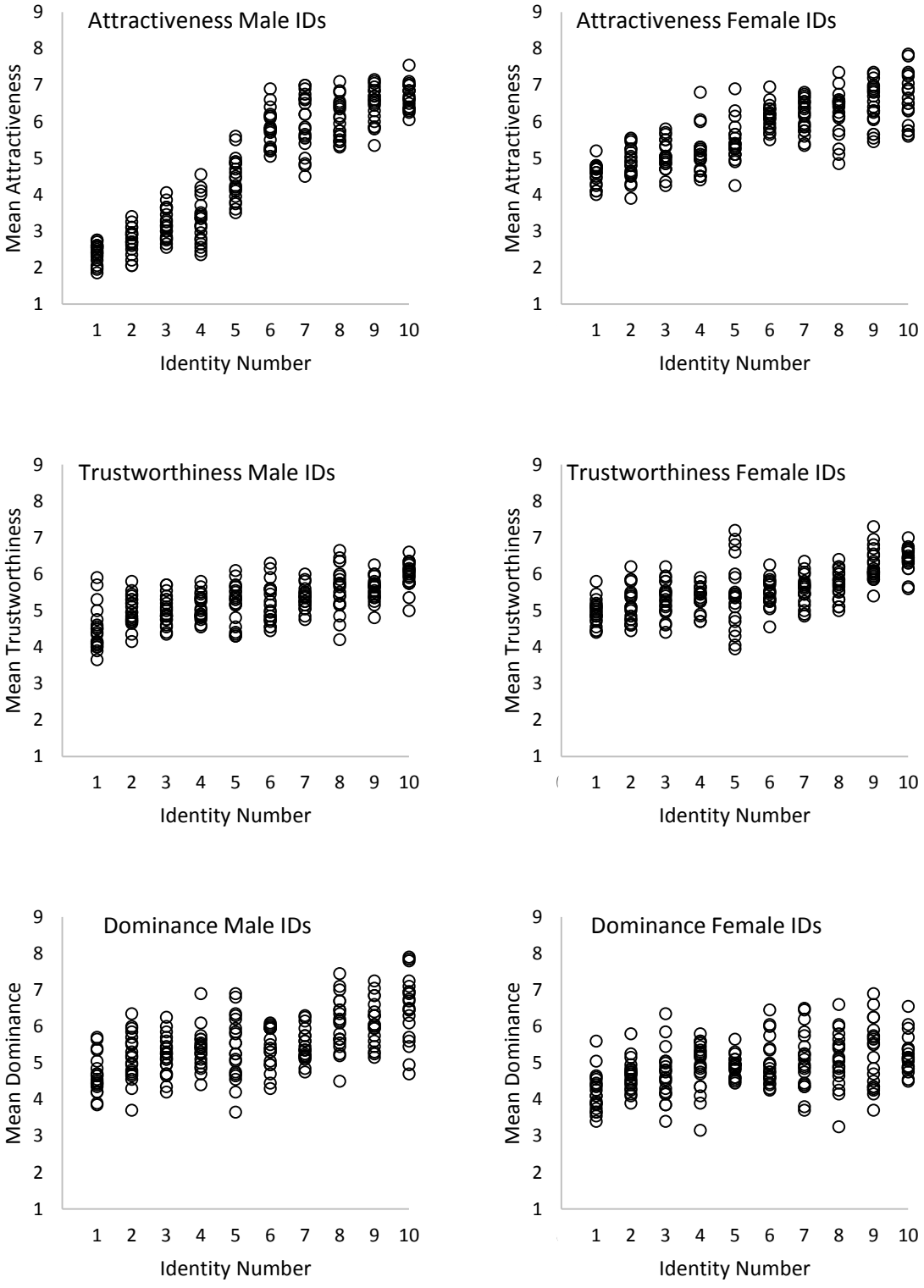

Figure 3.2. Mean ratings of all images from the 20-20 set for attractiveness (top), trustworthiness (middle) and dominance (bottom), displayed separately for male (left) and female (right) identities. Each column represents a single identity and each point represents a single image. Identities are ranked on the x-axis by mean identity score.

Table 3.1. *Variance in Social Attribute Judgements Between and Within Identities, Separately for Male and Female Identities.*

Attribute	Variance for male faces		Variance for female faces	
	Between IDs	Within IDs	Between IDs	Within IDs
Attractiveness	2.80	0.27	0.55	0.27
Trustworthiness	0.17	0.22	0.21	0.24
Dominance	0.29	0.42	0.10	0.24

These results provide strong evidence for the importance of within-person variation in social attributions of trustworthiness and dominance. In short, determination of these attributions relies just as heavily on choice of *photo* of a particular person, as it does on the *person* depicted in the photo.

Extracting and visualising within- and between-person variability associated with differences in social evaluation

Image processing

Prior to PCA images were scaled to 190 x 285 pixels and represented in RGB colour space using a lossless image format (bitmap). Face shape was derived by manually aligning the points of a standard grid with anatomical landmarks. The grid consisted of 82 xy-coordinates resulting in a shape-vector of 164 numbers (82 points x 2 coordinates) for each image. In order to derive texture-vectors the average shape across the whole image set was calculated. The texture for each image was then morphed to the average shape. This generated a texture-vector of pixel intensities comprising 162,450 numbers (190 width x 285 height x 3 RGB layers). PCA was performed separately for shape and texture. This generated a number of shape and texture eigenvectors and the original images were then recoded in the space

of these eigenvectors providing each image with a unique set of reconstruction coefficients of mean zero, which act as its signature. The projection of contributing faces onto the resulting eigenvectors is known as the 'reconstruction' of the face and we express this reconstruction in a low-dimensional space using the early eigenvectors of shape and texture. For the purposes of the present study, the first 30 principal components for shape and texture were used to model the variability of each identity and each image was coded as a set of 30 shape and 30 texture coefficients. Full details of this procedure can be found in Burton, Kramer, Ritchie, and Jenkins (2016).

Using image properties to predict attributions

High inter-rater reliability for the image ratings suggests that people use some consistent physical information in the face to inform their judgements. Here we ask whether a PCA image analysis can capture that physical information. We used the 60 derived dimensions from PCA (30 shape & 30 texture) to predict social attribute ratings using stepwise linear regression. The proportion of variance explained for each attribution across the whole set is shown in Table 3.2 (top row). We report adjusted R^2 rather than R^2 as a different number of PC components are extracted for each model from the stepwise regression. Adjusted R^2 accounts for this and makes the comparison between the models easier. Table 3.2 also shows proportion of variance accounted for when separate analyses are conducted for male and female face images (200 per set).

There are a number of interesting effects here. First, all social attributions are predicted to a high degree by image properties as captured by PCA (p values for all models < .001). This approach, therefore, provides an appropriate tool for exploring the properties of images that give rise to particular attributions. Second, the pattern among different attributions is interesting. Consistent with Todorov & Porter (2014), attractiveness is best captured by this analysis, with men's attractiveness being particularly well-predicted here (note that the large majority of raters were women). For male faces, dominance was better predicted than trustworthiness, a pattern which

reverses for female faces. This is consistent with research showing that, for men's faces, perception of dominance is very closely associated with perception of masculinity (Perrett et al., 1998). Since masculinity has clear biological markers in the face (Enlow & Hans, 1996), it seems reasonable that the image-level analysis will be able to pick these up. In contrast, perception of dominance in female faces is more complex (Keating, 1985) and may be less easy to predict from physical features, as shown in the relatively smaller proportion of variance accounted for in Table 3.2. This pattern of results is also consistent with general social evaluation models which argue that traits associated with the warmth component (equivalent to trustworthiness here) are particularly important for the evaluation of female faces, whereas traits associated with competence (equivalent to dominance here) are particularly important for the evaluation of male faces (Cuddy et al., 2008; Prentice & Carranza, 2002).

Table 3.2. *Using PCA to Predict Social Judgements Made to Ambient Images. Values Show Adjusted R^2 for an Analysis of all 400 Images (Top Row) and Separately for Males and Females (200 Images each).*

Identities	Attractiveness	Trustworthiness	Dominance
All	.72	.63	.59
Males	.83	.59	.65
Females	.64	.66	.46

Visualising predictors of social attributions

Dimensions derived from PCA can be used to build novel images. By manipulating the relative contributions of particular dimensions, this technique can be used to visualise the aspects of faces which these dimensions code. Figure 3.3 shows examples of this manipulation, using the

60 derived dimensions (30 shape & 30 texture) from the analysis of the entire set described above.

For two individuals, we have first reconstructed a particular image in PC-space (top row of Figure 3.3). This illustrates that the PCA captures much of the relevant variance in the original 400-image set. The original images are reproduced rather well in this low-dimensional representation (i.e. using just 60 coefficients, rather that the many thousands necessary to represent the original images pixel-by-pixel). So, the reconstructions of the two faces differ only in the relative weightings of the 60 PCA dimensions, but nevertheless the reconstructions seem to preserve much of the character of their corresponding originals.

We next adjust these reconstructed images in a manner intended to render them high or low in the three social attributions used in this study (rows 2 and 3 of Figure 3.3). For each rated attribute (attractiveness, trustworthiness, dominance), we take those PC dimensions which significantly predict viewers' ratings in the analysis above. We then reconstruct the target face (row 1 of Figure 3.3) such that predictive components are given a value of ±1z (i.e. one standard deviation of the distribution of that component across the whole 400 images, weighted positive or negative according the direction of correlation). All other components retain their original value. This technique tends to emphasise dimensions predictive of a particular social trait, without giving prominence to any individual dimension.

This visualisation technique appears to be promising. The manipulated images give plausible dimensions, in that high and low variants on each attribute appear to result in corresponding face images. In particular, note that trustworthiness and dominance manipulations also affect expression in ways consistent with well-established associations – i.e. smiling faces are seen as more trustworthy and less dominant than unsmiling faces (Montepare & Dobish, 2003; Said et al., 2009). These visualisations also emphasise the characteristics of the original set. Note that each manipulation

also appears to affect the *identity* of the face. Because this set includes 20 images of 20 people, changes in trustworthiness (say) result in changes between apparently trustworthy and untrustworthy people, as well as changes common to everyone, such as expression. In our next study, we focus on *within-person* variability and explore whether it is sufficient to produce similar changes in social perception without changes in identity.

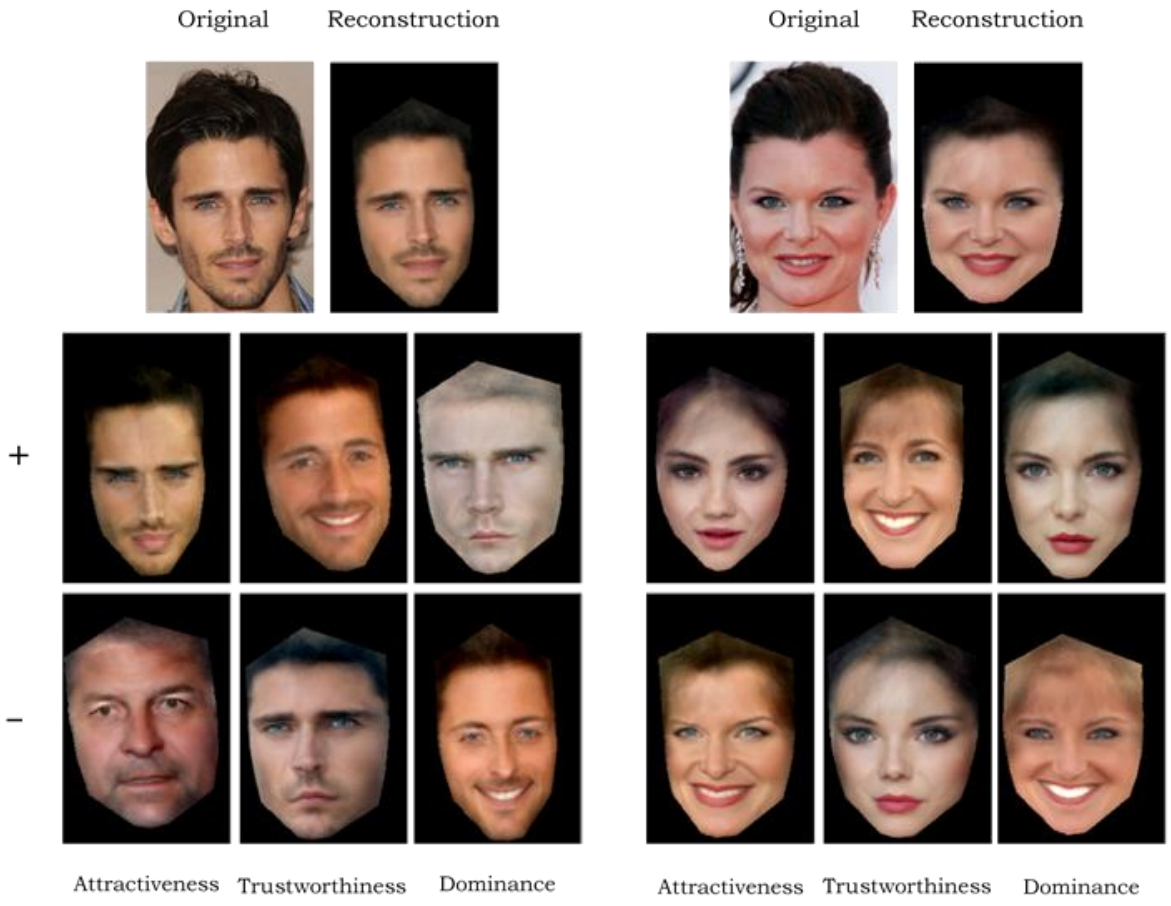

Figure 3.3. Two example images reconstructed from 60 PCA components (top row), and reconstructed to emphasise dimensions predicting social traits (rows 2 and 3) using the 20-20 image set.

3.3 Experiment 4

Introduction

In Experiment 4 we examined the range of social attribute ratings within photos of the same person. The intention was to gather a large sample of images for each person, which could be used to derive an idiosyncratic space of variability representing only that person. We can then use properties of an individual's space to predict attributions made to photos of that person. In order to do this, we carried out analyses similar to those in Experiment 3, but this time using 100 images of each of four individuals. We gathered social ratings for each person and examined the variance of these by individual and social attribution. We then carried out PCA on images separately for each person, allowing us to extract the variation leading to social attributions *for that person*. Research in facial *identity* has shown large idiosyncratic differences in the type of variance different faces exhibit (Burton et al, 2016) and this has been used to explain differences in familiar and unfamiliar face recognition (Burton, 2013; Young & Burton, in press). Here we ask whether these differences give rise to idiosyncratic social attribution.

This novel approach allows us to eliminate any identity-based differences and focus on the magnitude of social variance produced purely by the statistical properties of images. Once idiosyncratic information relevant to each social attribute is extracted, we can use it to change the way images of these identities are perceived by manipulating their properties while preserving identity-diagnostic information. Results are consistent with findings from Experiment 3 in showing large identity-based differences for attractiveness attribution but larger differences within each identity for trustworthiness and dominance attribution. Using the models to visualise information relevant to each social dimension for each identity also demonstrates that the extracted within-person variability is sufficient to bring about significant and meaningful changes in social evaluation.

Method

Participants

Images were rated by 40 participants (9 male, mean age = 20.5, age range = 18-25), all from the University of York. All had normal or corrected-

to-normal vision and received payment or course credit for their participation. Participants provided informed consent prior to their participation in accordance with the ethical standards stated in the 1964 Declaration of Helsinki. Experimental procedures were also approved by the ethics committee of the University of York psychology department.

Stimuli

Stimuli were 100 images of each of four individuals (2 male), all of which were identities from the 20-20 set. These were foreign celebrities and associates, unfamiliar to UK viewers. Selection criteria were the same as in Experiment 3: images were downloaded from internet search on names, and the first 100 images returned were chosen for which faces were not obscured by clothing or glasses.

Results and discussion

Ratings

All 400 images were rated for attractiveness, trustworthiness, and dominance on a nine-point scale. Images were presented in a separate random order for each participant, and responses were given via a mouse click. Each participant rated 50 images per identity (200 in total) and all images were rated by 20 participants.

As in Experiment 3, there was very high inter-rater reliability for each of the scales (Cronbach's alphas above .88 for all three social attributes, Nunnally, 1978). Figure 3.4 shows the spread of ratings for each social trait across all four identities, which we label M1, M2 (males) and F1, F2 (females). Consistent with results from Experiment 3, there are large between-person differences in ratings of attractiveness, but not in ratings of trustworthiness or dominance. Our intention in this analysis is not to make formal between-person comparisons, but it is nevertheless interesting to observe that the patterns of ratings are very similar to the initial study: two of the identities

are consistently rated as more attractive than the other two. In contrast, judgements of trustworthiness and dominance are highly overlapping.

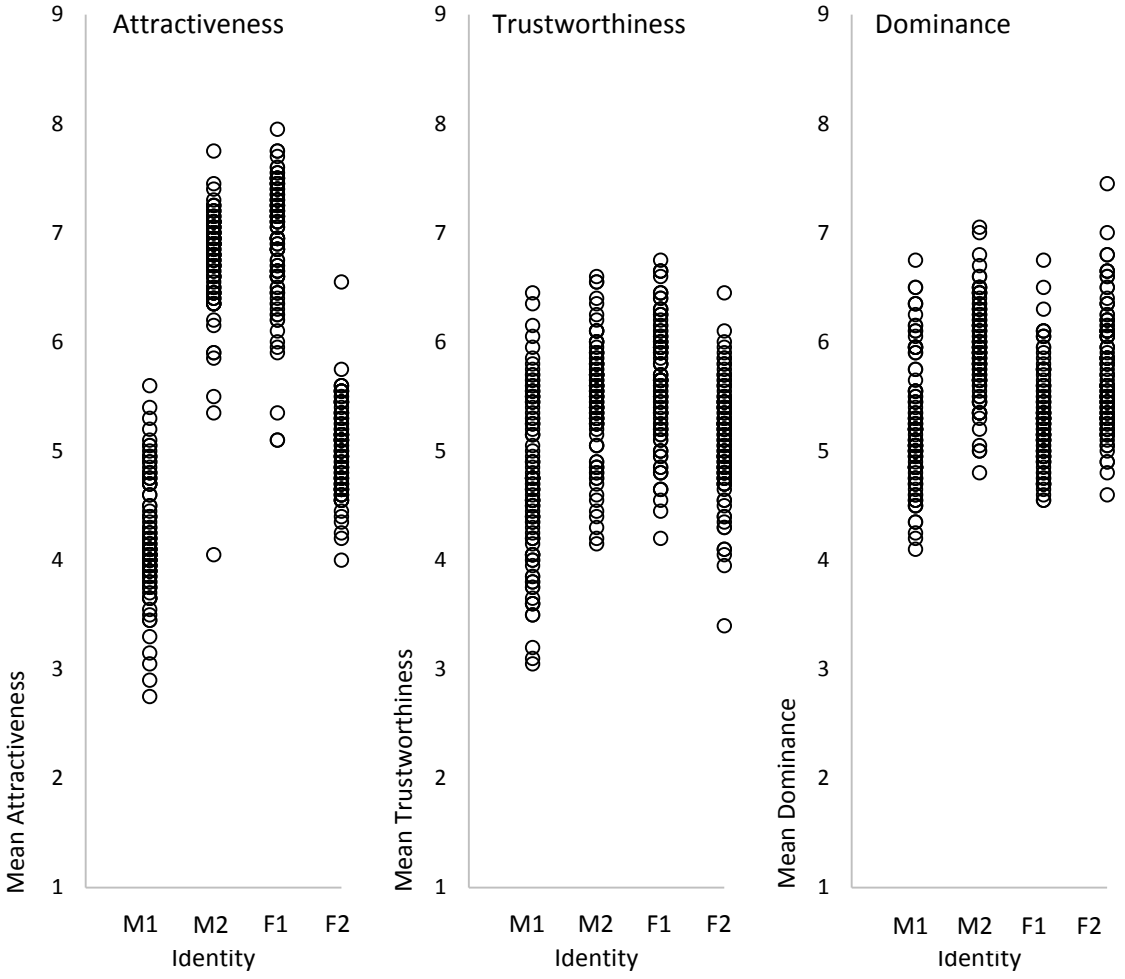

Figure 3.4. Mean ratings of 100 images for each of four people (males: M1 and M2, females: F1 and F2). Ratings are shown for attractiveness (left), trustworthiness (middle) and dominance (right) for each identity.

Image analysis

We next conducted within-person PCA on the 100 images of each target person separately. Our intention here is to establish, separately for each person, how well the variance in social attributes they receive is predicted by the image properties of their own photos. This differs from Experiment 3, in which predictive dimensions were derived from an image set

containing both within- and between-person variations (20 images of 20 people). Instead, the current analysis is conducted person by person.

For each person, we derived 60 components (30 shape & 30 texture) from a PCA on 100 images. We then used these 60 components as predictor variables in a stepwise linear regression analysis taking human ratings as the dependent variable. Table 3.3 shows the proportion of variance captured for each rating scale, separately for each of the four identities. Again, we used R^2 adjusted rather than R^2 in order to allow comparisons across the different models.

Table 3.3. *Variance Explained (R2 adj.) in Predicting Social Judgements for each of Four Identities (Male IDs: M1 & M2, Female IDs: F1 & F2)*

Identity	Attractiveness	Trustworthiness	Dominance
M1	.38	.63	.54
M2	.66	.30	.58
F1	.37	.65	.65
F2	.46	.63	.64

Results show rather idiosyncratic patterns in the levels to which image analysis predicts social judgements. All models explain a significant amount of variance with all p values < .01, demonstrating that differences in the statistical properties of images can account for at least some variance in social attribution. Across identities, most variance is explained for dominance, followed by trustworthiness and then attractiveness. Across traits, however, the pattern of predictive power is different for each of the identities. Focusing on the amount of variability explained for each identity separately, it is clear that some models are more successful than others. Identities have their own idiosyncratic ways of varying for each trait, suggesting that cross-person investigations of the relationship between traits and physical properties do not tell the whole story.

The trustworthiness model for M2 is of particular interest, as it presents somewhat inconsistent results compared to both the variance explained for attractiveness and dominance for that specific identity and the trustworthiness variance explained for the other three identities. The range of trustworthiness ratings presented in Figure 3.4 shows comparable spread across all identities, eliminating an explanation in terms of limited range. Visual inspection of the images rated as the most and least trustworthy for this specific identity provides one possible explanation. Figure 3.5 shows the three images rated as most and least trustworthy for the two male identities. Images for M1 show a clear association between emotional expression and trustworthiness – i.e. smiling faces are judged as trustworthy. This fits well with earlier population-level research and is explained by the emotion overgeneralisation hypothesis (Said et al., 2009; Zebrowitz et al., 2010). However, this general pattern does not hold for M2 – for whom emotional expression does not predict trustworthiness ratings. This is particularly interesting, because it shows how associations derived from one image per identity (as with most research), give rise to associations which do not generalise to all people. From the results in the literature, one might decide to choose a smiling photo to present oneself as trustworthy, as that is the general finding. However, this association is clearly not true for this individual, M2 (Perhaps he has an odd smile, or one that makes him look sinister?). Only within-person analyses can reveal such associations.

Visualising within-person predictors of social attributions

As with Experiment 3, the dimensions derived from PCA were used to reconstruct the original images and manipulate them to alter social evaluation. The key difference here is that as PCA was applied separately for each identity, it extracted idiosyncratic information only. Images presented in Figure 3.6 show manipulations of the same two base-images as in Figure 3.3. For each identity, the original image was first reconstructed using 30 shape and 30 texture components derived from that person's specific PCA (based on 100 images). In order to manipulate the perception of each image we then identified the specific PC dimensions predicting a significant amount of variance in attribute ratings, and assigned a value of ±1z to them depending

on the direction of correlation. The values of the PCs which were not associated with differences in social attribution remained unchanged. This procedure was used separately for each trait and identity; therefore, only idiosyncratic, identity-specific information was used to change the way these images are perceived.

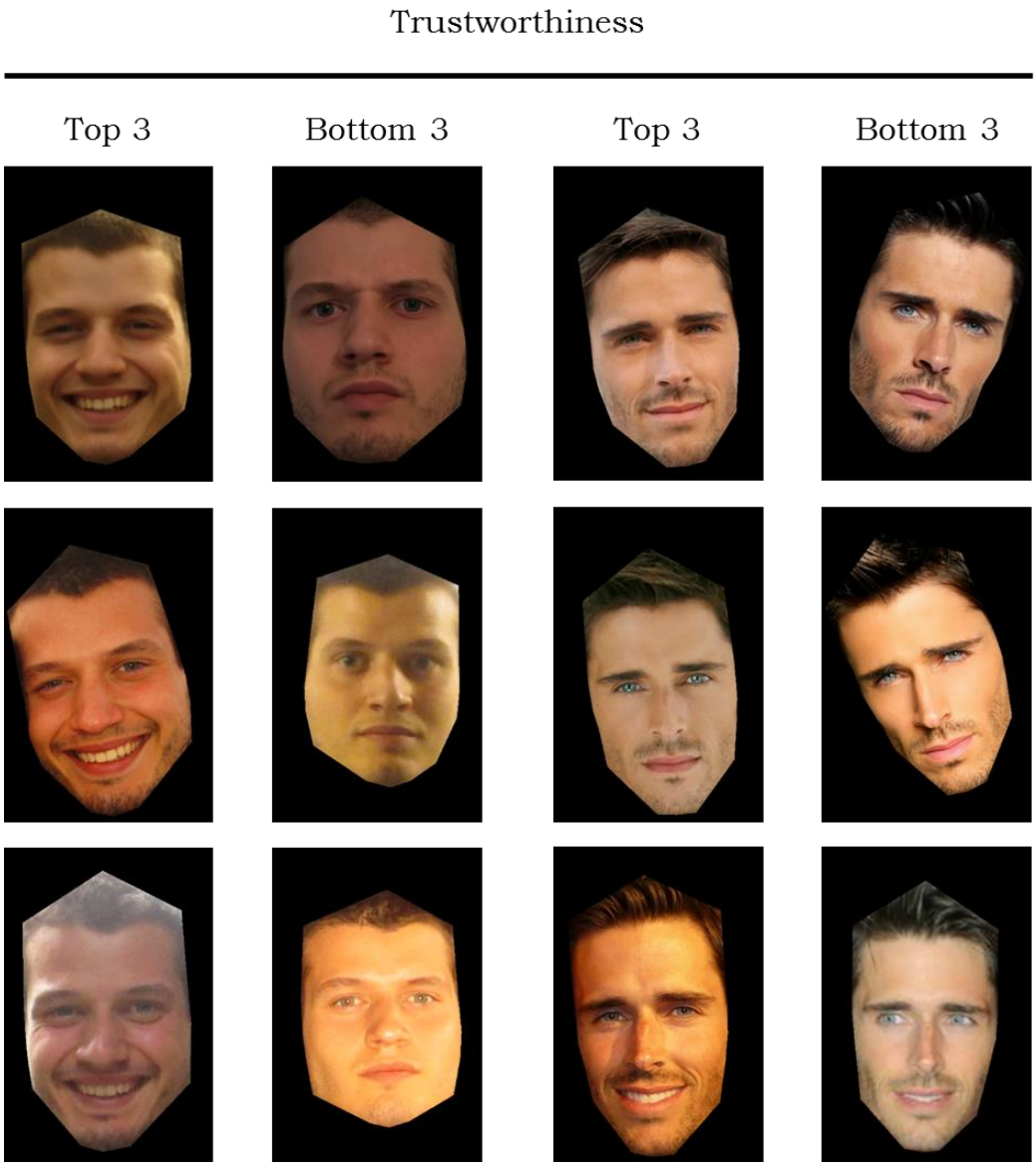

Figure 3.5. Images rated as the most and least trustworthy for M1 (left) and M2 (right).

Reconstructed images in Figure 3.6 (top row) show high resemblance to the original photos, demonstrating that the low-dimensional PC space is a good representation of these images. More importantly, manipulated images

appear to produce the predicted changes: the faces seem to look more or less attractive, trustworthy, and dominant, consistent with the image manipulations. Comparing these images to Figure 3.3 highlights some similarities, especially in the importance of emotional expressions. Again, smiling faces are perceived as more trustworthy and less dominant which fits well with the emotion overgeneralisation hypothesis (Montepare & Dobish, 2003; Said et al., 2009). The critical difference between the two approaches, however, is identity. So, what we have manipulated here is the structure that makes each *person* look more or less trustworthy etc. In order to support these impressions, in the following two studies we present manipulated images such as those in Figure 3.6, and seek viewers' real social attributions to them.

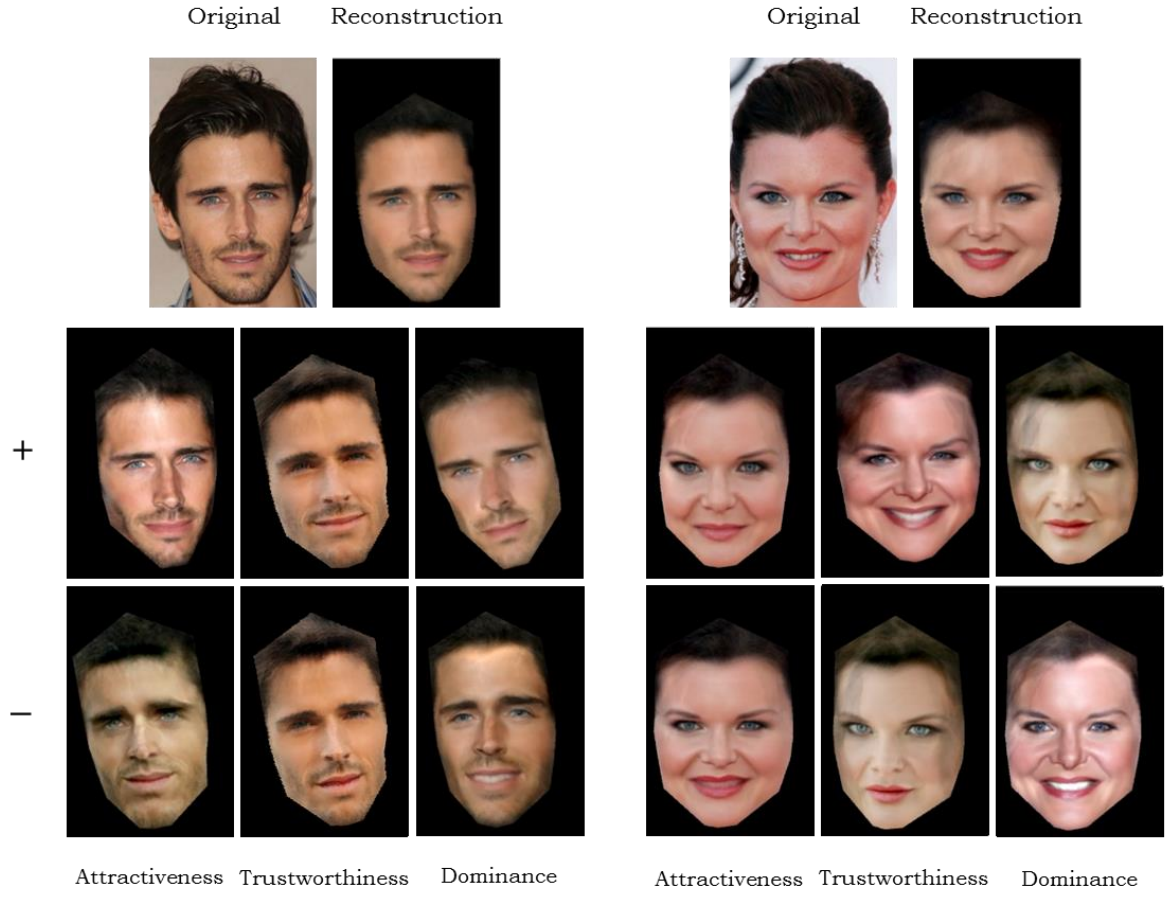

Figure 3.6. Two example images reconstructed from 60 PCA components (top row), and reconstructed to emphasise dimensions predicting social ratings (rows 2 and 3) using the within-person image set. To make visual comparison easier, these are the same identities and images as in Figure 3.3.

3.4 Experiment 5

Introduction

Experiments 5 and 6 aim to validate our approach for investigating the physical face dimensions people rely on when evaluating face images. Here we use the statistical components derived for each face in Experiment 4 to manipulate images in ways predicted to affect social judgements. The validity of these manipulations is established by asking viewers to discriminate between pairs of faces in judgements of attractiveness, trustworthiness, and dominance. In Experiment 5, we manipulate images from the original corpus, i.e. those which were used to derive the statistical description of each person. In Experiment 6, we apply the derived dimensions to entirely new images of the same people. To anticipate the results, in both experiments we find clear correspondence between perceived and manipulated dimensions of trustworthiness and dominance. However, there is a much less clear association for judgements of attractiveness.

Method

Participants

A total of 80 participants (20 male, M = 21.4 years, age range: 18-27) from the University of York took part in the study. All had normal or corrected-to-normal vision and received payment or course credit for their participation. Participants provided informed consent to experimental procedures, which were approved by the ethics committee of the University of York psychology department. Only participants who had not taken part in Experiment 4 were recruited for the present study.

Stimuli

Forty images of each of the four identities from Experiment 4 were used to create a total of 480 image pairs (each image was manipulated for all three social attributes). Each image pair contained reconstructions of the same

base image with one image modified towards a higher and the other towards a lower degree of the same attribute by changing the value of the shape and texture components correlating with this attribute to ±1z, depending on the direction of the correlation. Shape and texture components which accounted for the variance in social attribute ratings were used in combination to create these new manipulated images. See Figure 3.7 for examples.

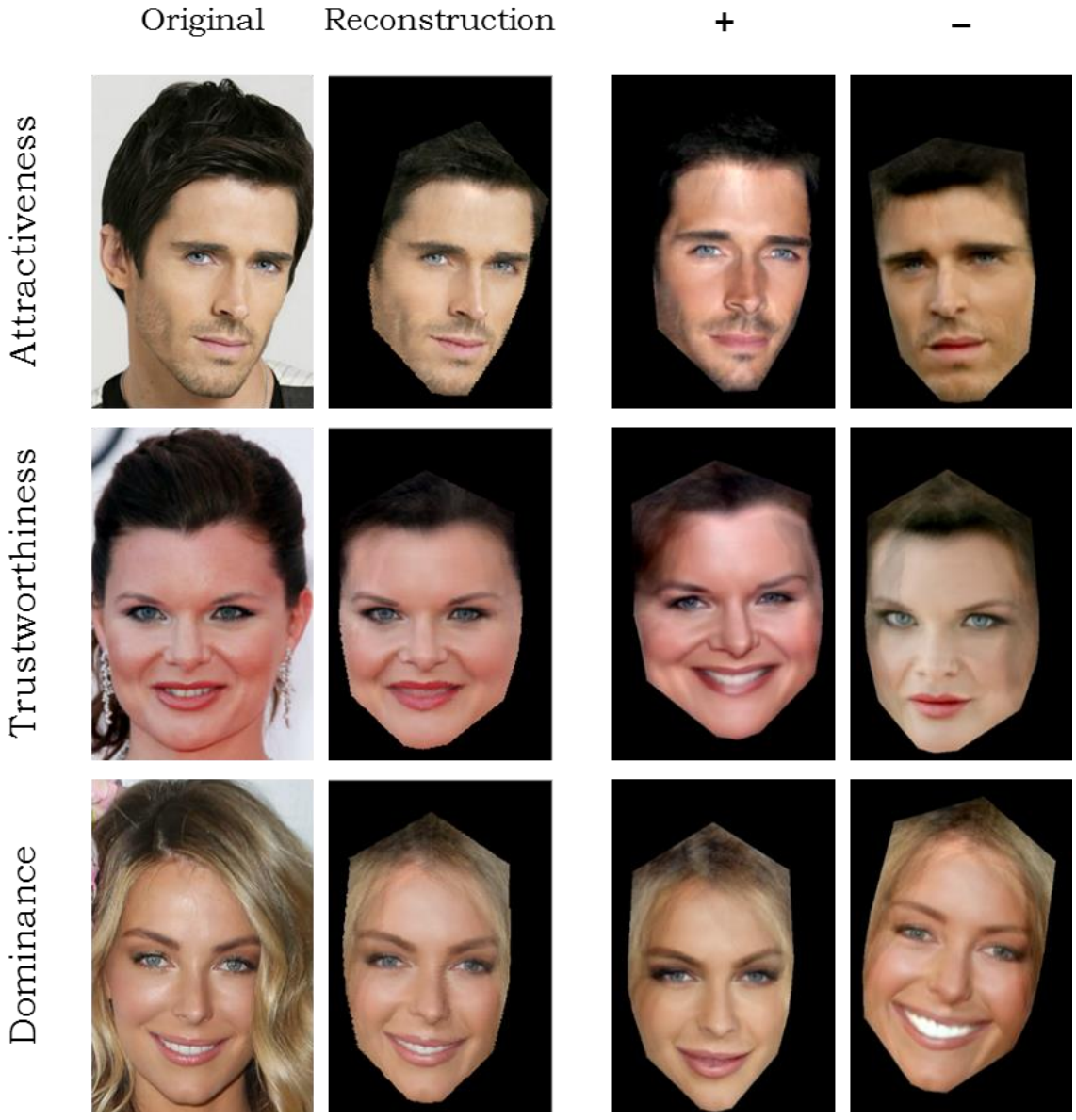

Figure 3.7. Examples of image reconstructions and manipulated pairs used as stimuli in Experiment 5.

Design and procedure

Each participant completed 120 trials, in which they were asked to judge which of a pair of images was more attractive, trustworthy, or dominant. Responses were made by key-press. Each image in a pair depicted the same person, manipulated to predict high or low attributions of that dimension, as in Figure 3.7 (pairs of images of the right half of the figure). Trials were blocked by rating dimensions, i.e. 40 trials for each of attractiveness, trustworthiness, and dominance, and blocks were separated by a short rest. The order of block and image pair presentation was randomised per participant. There were four versions of the task, each representing a quarter of the available image pairs. These were counterbalanced across the experiment, so that each image pair was seen by only 20 participants.

Results and discussion

In order to test whether image manipulation successfully reflected the social perception of the four identities, the mean proportion of manipulation-consistent responses was calculated for each identity and each trait (see Figure 3.8). As participants were presented with an image pair for each trial and asked to identify the image that was more attractive, trustworthy, or dominant their responses indicated whether they were able to detect the direction of our manipulation. One-sample t-tests against chance (50%) were performed for each identity (one each for attractiveness, trustworthiness, and dominance) with alpha levels corrected for multiple tests. The intended direction of our manipulation was detected significantly above chance for all manipulations (t_{min} (79) = 2.23, p_{max} = .029), except for the attractiveness ratings to person F1 which were significantly counter to predicted direction, t (79) = 4.48, p < .001, d = 0.50. Cohen's d statistic (from 0.90 to > 0.99) indicated large effect sizes across most identities and social traits, except for the person F2 attractiveness manipulation where Cohen's d showed a small effect (d = .25)

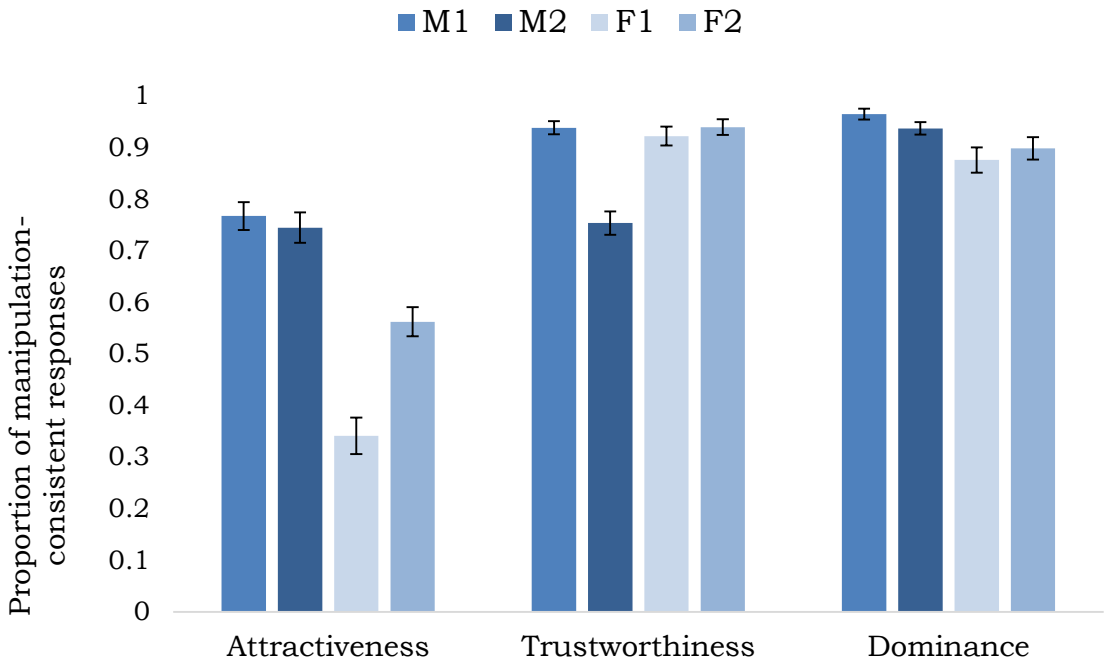

Figure 3.8. Mean proportion of manipulation-consistent responses for all identities and social attributes. High values indicate that participants were successful in identifying the directions in which the images were manipulated. Error bars represent standard error of the mean.

These results show that we have successfully captured the underlying physical information people use to inform their dominance and trustworthiness judgements and that within-person variability alone is sufficient to bring about meaningful changes in social evaluation. Highest accuracy was achieved for the dominance manipulation, which fits well with the results of the PCA and regression analysis reported earlier. Moreover, manipulation consistency across all identities was higher for trustworthiness than attractiveness and this is also well reflected in the results of the regression analysis in Experiment 4.

A surprising finding was the low detection rate for the attractiveness manipulation for identity F1. Visual inspection of the images rated as most and least attractive for this particular identity indicated that participants consistently rated images with good quality as more attractive than images with poor quality (sharper and more heavily pixelated images). As our image reconstruction process inevitably smooths over and blurs the original images

these quality differences are lost - which could explain the low detection rate for this specific trait and identity. Overall, and with this single exception, results show that image manipulation was successful and that within-person variability predicts attributions.

3.5 Experiment 6

Introduction

The next validation study examines whether we have extracted sufficient within-person information to generalise to completely novel images. Ten new images were collected for each of the identities used in Experiments 4 and 5, and we projected these into the identity-specific PCA-spaces already derived above. As with Experiment 5, we manipulated each of the novel images in ways predicted to increase or decrease attributions of attractiveness, trustworthiness, or dominance.

Method

Participants

A total of 75 participants (20 male, M = 21.4 years, age range: 18-27) from the University of York took part in the study. All had normal or corrected-to-normal vision and received payment or course credit for their participation. Participants provided informed consent to experimental procedures, which were approved by the ethics committee of the University of York psychology department and only those who had not taken part in the previous studies were recruited.

Stimuli

Ten new images were collected for each identity (total of 40 images) and used to create image pairs. Each new image was landmarked in the same way as the original images and reconstructed using the components from the original 100 images for that particular identity. Each image was then

manipulated to look more or less attractive, trustworthy, or dominant, in exactly the same way as described for Experiment 5. Figure 3.9 shows examples.

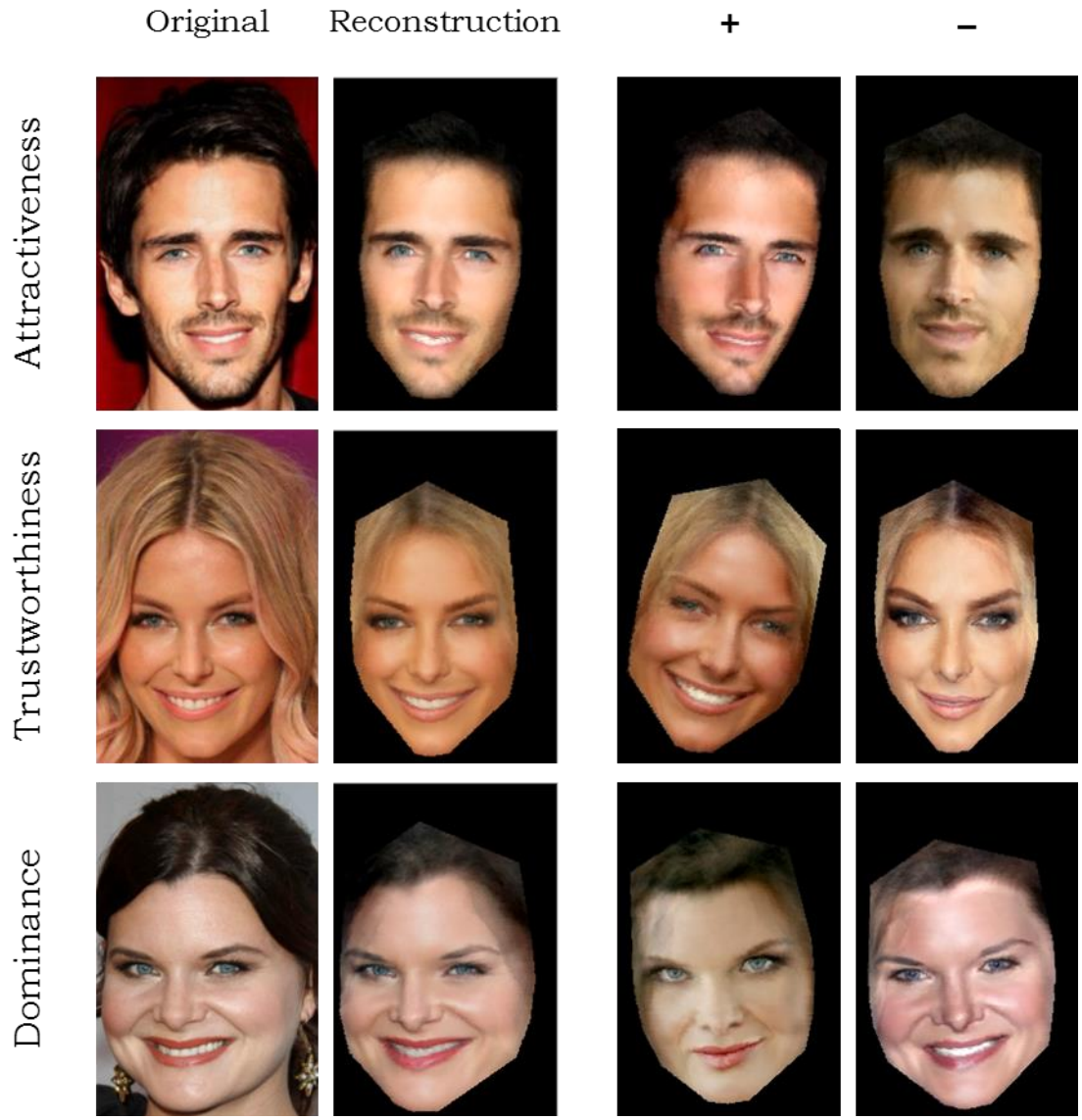

Figure 3.9. Examples of image reconstructions and manipulated pairs used as stimuli in Experiment 6.

Design and procedure

This study followed the same design and procedure as Experiment 5. Participants were presented with an image pair on the screen and asked to decide which image looked more attractive, dominant, or trustworthy. Prior to

the experiment participants were told that the study was about first impressions from faces and that they would see pairs of similar face images. Attribute decisions were blocked and stimuli were randomised within blocks, in the same way as for Experiment 5.

Results and discussion

Mean proportion of manipulation-consistent responses was calculated for each identity and social trait (see Figure 3.10). Comparison with chance (50%) showed that most judgements were significantly consistent with the manipulation. However, just as in Experiment 5, attributions of attractiveness made to person F1 ran significantly counter to manipulation, $t(25) = 2.81$, $p < .05$, $d = 0.56$. Furthermore, judgments of attractiveness for person F2 were not significantly different from chance levels, $t(25) = .87$, $p > .05$, $d = 0.17$. All other judgements showed significant manipulation-consistent responses (one sample t-tests, $t_{min}(25) = 2.34$, $p_{max} = .028$). Cohen's d statistic (from 0.47 to > 0.99) indicated medium to large effect sizes for all other identities and social traits.

Results again demonstrate that, overall, participants were able to identify the direction of attribute manipulation. The pattern of results is very similar to the one for the original images with highest accuracy for dominance and lower accuracy for attractiveness in the female identities. Altogether, results from both validation studies demonstrate that within-person variability is large enough for us to be able to manipulate it in a meaningful way. In particular, it seems consistently possible to manipulate judgements of trustworthiness and dominance – though attractiveness is less consistent. More importantly, changes in social evaluation do not require changes in identity, highlighting the importance of the statistical properties of images.

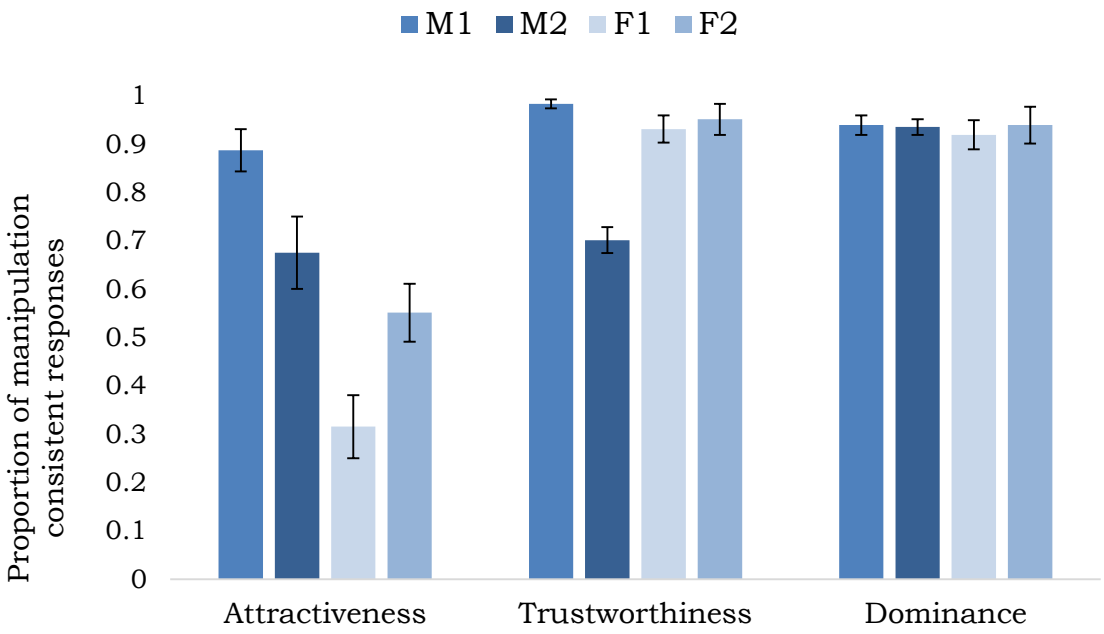

Figure 3.10. Mean proportion of manipulation-consistent responses across social attributes for the novel images of each identity. High values indicate that participants were successful in identifying the directions in which the images were manipulated. Error bars represent standard error of the mean.

3.6 Experiment 7

Introduction

The last experiment in this chapter outlines a more exploratory analysis of the low-level image properties and face metrics that could have a significant influence on social face evaluation. These image measures include colour and texture estimates such as contrast, brightness, and the amount of redness in the face as well as shape-related measures such as facial width-to-height ratio. Differences in textural components have been mostly associated with the evaluation of attractiveness. Perrett and colleagues, for example, argue that an increased amount of redness in the face signals better health and therefore will lead to higher ratings of attractiveness (Stephen et al., 2011; Stephen & Perrett, 2015). This association has been supported by studies both measuring and manipulating facial redness (Pazda et al., 2016; Stephen et al., 2009). Contrast is another low-level measure linked to attractiveness perception. It has been mostly related to the use of facial

cosmetics which usually enhance facial contrast and thus boost ratings of attractiveness (Porcheron et al., 2013). Similarly, there is evidence that images with increased brightness are also evaluated in a more favourable way (Lakens et al., 2013). In addition to attractiveness, an increase in both redness and contrast has been associated with higher ratings of dominance (Stephen et al., 2012), although the effect of contrast seems to be limited to female participants rating female faces (Mileva et al., 2016). Furthermore, measures more commonly used in natural scene perception have also been shown to influence attractiveness evaluation. Analyses of large scale data sets of natural scenes and artworks show that images with a Fourier slope closer to -2 are generally perceived as more aesthetically pleasing (Graham & Field, 2007; Redies et al., 2007). Relating this to first impressions from faces, Menzel et al. (2015) demonstrated a similar pattern of results where face images with a shallower slope (closer to -2) were also rated as more attractive than images with a steeper slope.

Facial width-to-height ration (FWHR) is the shape measure that has received the most research attention in social evaluation. It is commonly associated with the attribution of dominance and aggression, with wider faces perceived as more dominant than longer faces (Carre & McCormick, 2008; Carre et al., 2009). Stirrat and Perrett (2010) further report that wider faces are also perceived as less trustworthy. Nevertheless, some argue that there are large within-identity differences for this metric as it fluctuates significantly when displaying different emotional expressions (Kramer, 2016). This is particularly relevant here as results from Experiment 4 suggest that social evaluation depends on these physical image properties even when rating images of the same person. Therefore, Experiment 7 aims to explore the extent to which these measures can influence ratings of different images of the same person and whether the effects of these properties are stable across all identities or are rather idiosyncratic. The analysis included measures that have already been linked to first impressions such as brightness, contrast, redness, and FWHR as well as other low-level image measures which could be relevant to social evaluation, such as blur, face size and skin tone.

Method

Materials

All images from the within person image set (from Experiment 4) were used in the present experiment. These were 100 images each of four unfamiliar identities (2 male). All images were ambient and captured a great amount of everyday variability which allowed us to measure a range of image properties. All original images were represented in RGB colour space using a lossless image format.

Image measures

A variety of physical image measures addressing colour, texture, and shape differences were used as possible predictors of social evaluation. Calculations for all physical measures were performed using MATLAB. In order to avoid the influence of the image background, all colour and texture measures were taken within the face only. This was done by using the anatomical points describing the contours of the face from the PCA analysis in Experiment 4 (see numbered points in Figure 3.11). This section provides descriptions and definitions of each measure included in the analyses. Figure 3.12 provides examples of the images with the highest and the lowest values on all image metrics.

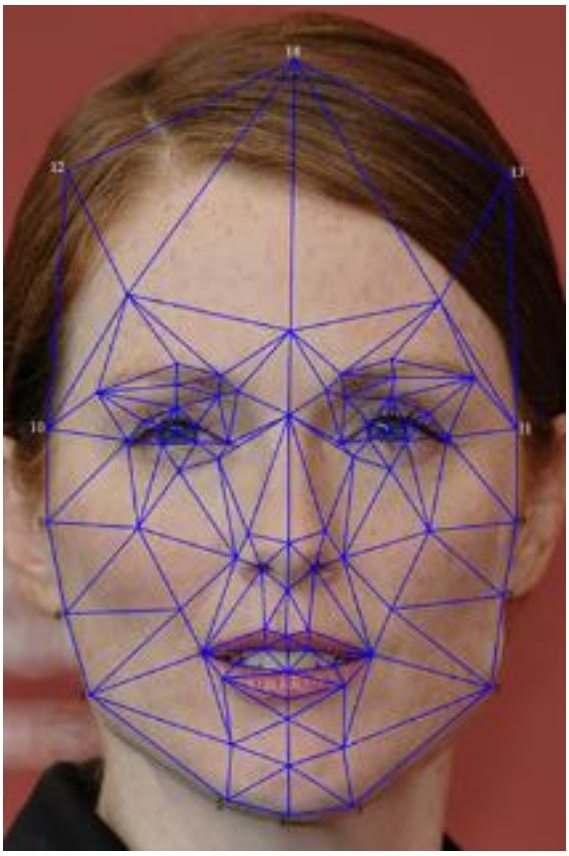

Figure 3.11. Landmarks layout.

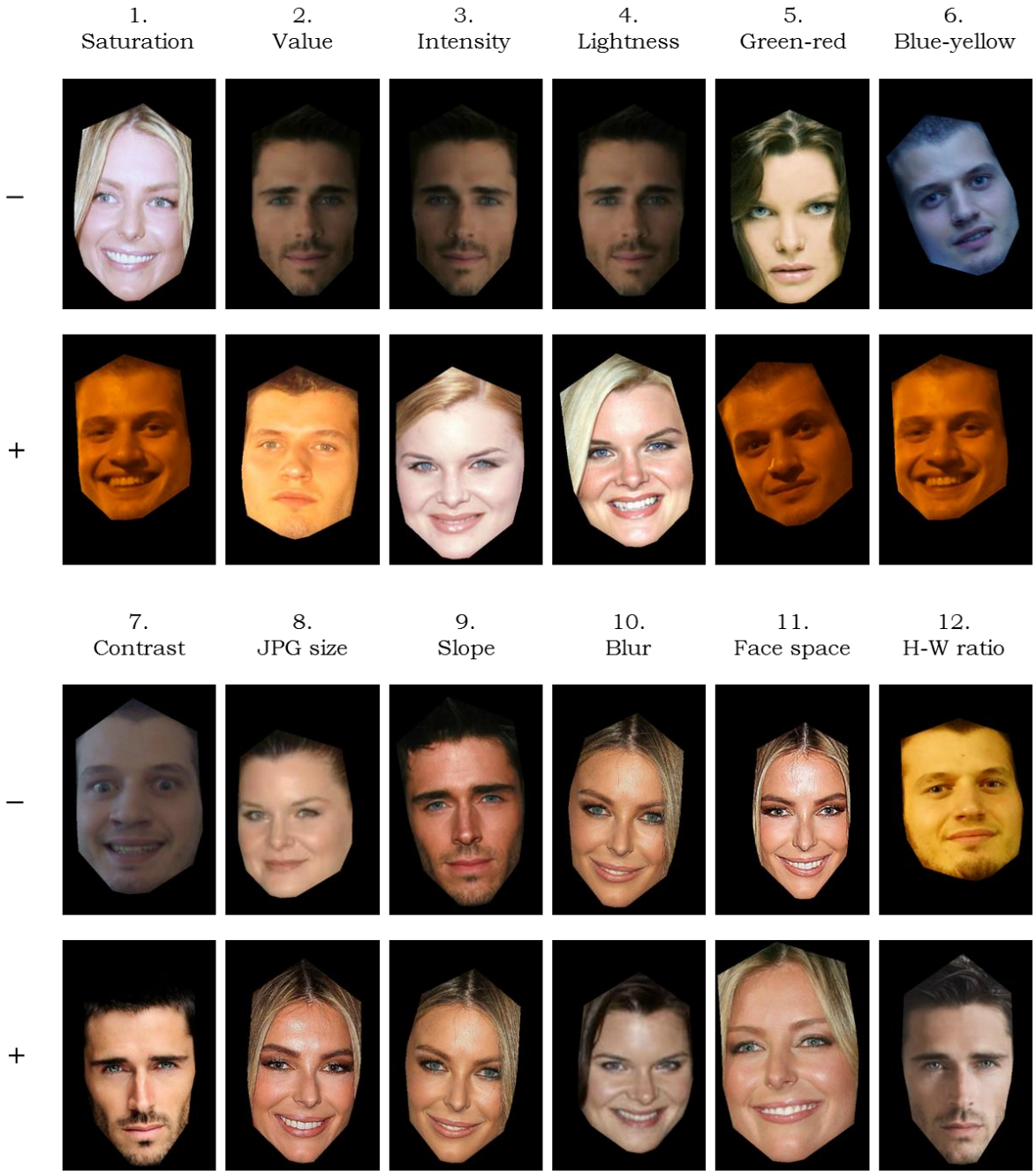

Figure 3.12. Examples of images measured as high and low on all physical measures included in Experiment 7.

1. *Saturation*

   In order to extract saturation information all images were transformed from RGB to HSV colour space (see bottom right on Figure 3.13). This space uses cylindrical coordinates to represent RGB points (Wen & Chou, 2004). The first two components in this space – H and S

represent hue and saturation respectively and depend on the human colour spectrum perception (Plataniotis & Venetsanopoulos, 2013). Hue relates to the specific colour perceived, whereas saturation measures the amount of white embedded in the specific hue. Thus, saturation represents the ratio of colourfulness to brightness in the image.

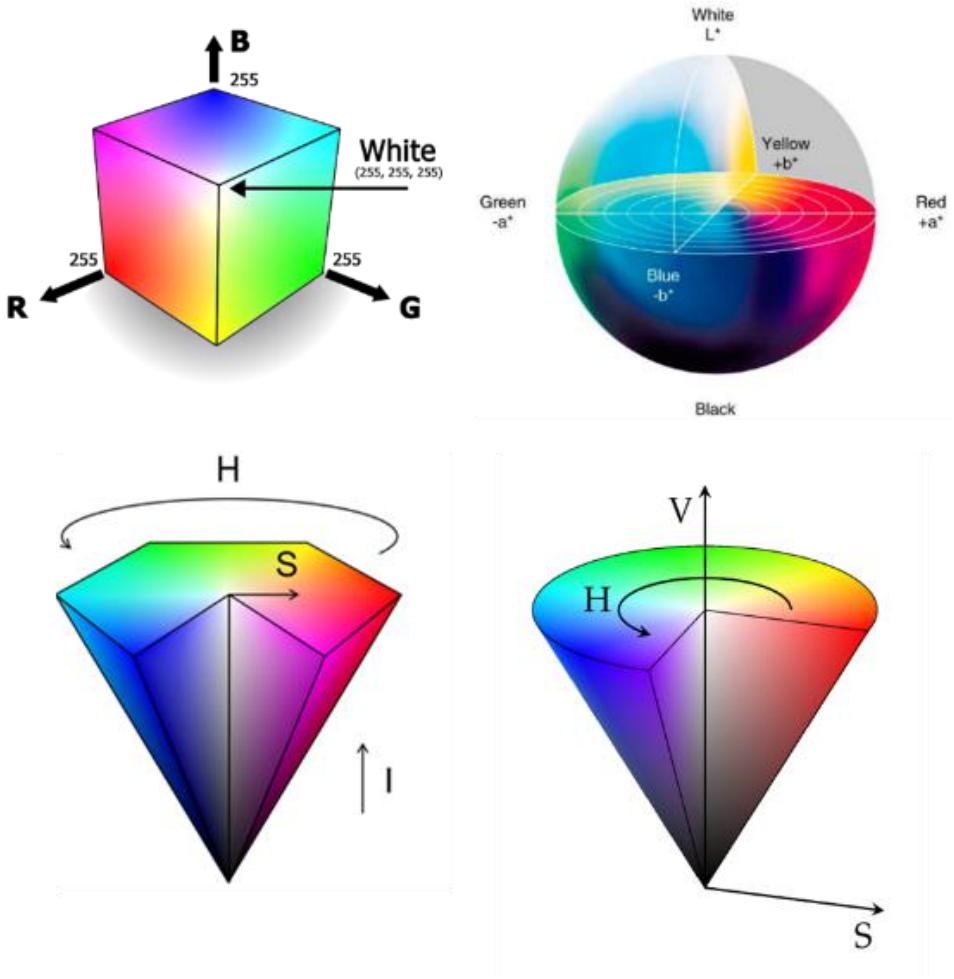

Figure 3.13. Different colour spaces used to measure colour differences in the face images. Top left shows the RGB space which was used to represent the original images. Top right shows the CIE Lab space, bottom left shows the HSI space and bottom right shows the HSV space.

2. *Value*

The same "hexacone" colour space (HSV) was used to measure the value of the images. Value is the last component of the HSV space and is defined as the largest component of a colour. This makes it a good

measure of brightness of colour. Figure 3.14b shows value plotted against chroma (colourfulness) for a pair of complementary hues and the formula used to calculate it.

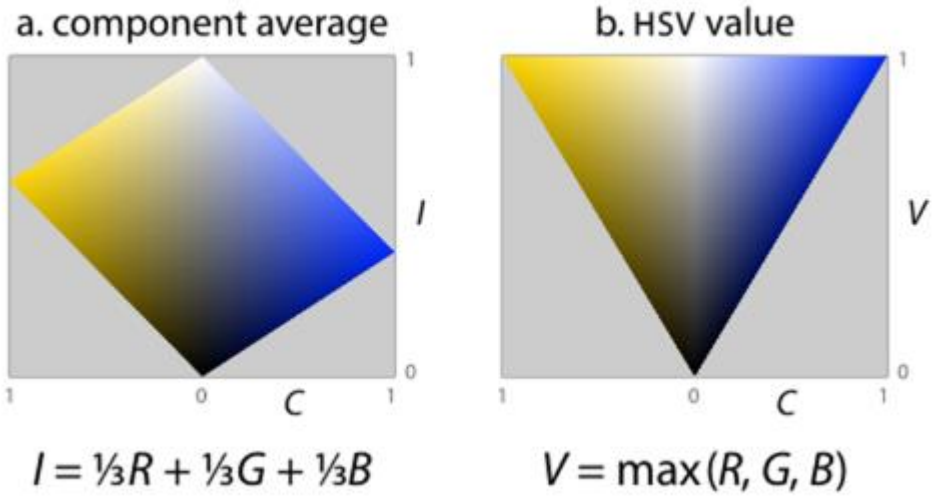

Figure 3.14. Difference between value and intensity as measures of image brightness and the formulae used to calculate them.

3. *Intensity*

   In order to extract intensity, all images were transformed from RGB to the HSI colour space (see bottom left on Figure 3.13). Just like HSV, this colour space is a linear transformation of the RGB space and uses cylindrical coordinates to represent RGB points (Wen & Chou, 2004). Intensity is the last component of the HSI space and it represents the average of the R, G and B components. Both value and intensity measure the brightness of the image, however, value measures the maximum of the three RGB channels, whereas intensity takes their average. This has very subtle implications. As can be seen in Figure 3.14 the highest values in the intensity dimension are associated with white specifically, whereas the highest points in the value dimension are associated with the brightness of each particular colour. Therefore, higher intensity refers to white while higher value encompasses different hues.

4. *Lightness*

   Another measure of brightness or lightness of colour comes from the CIE $L^*a^*b^*$ colour space (see top right in Figure 3.13). This space represents all perceivable colours mathematically, which makes it superior to the RGB space, as is only represents 90% of all perceivable colours. It is also device-independent, meaning that it represents colour without taking the nature of its creation or the device it is displayed on in consideration. In this space $L^* = 0$ refers to black and $L^* = 100$ refers to white.

5. *Green-red*

   The second component of the CIE $L^*a^*b^*$ colour space, a^*, was used to measure the amount of red in the image. This measure correlates with red-green chroma perception, with higher values of a^* indicating red and lower values indicating green. This dimension of the $L^*a^*b^*$ colour space has been used to measure as well as manipulate redness in the face in the attractiveness perception literature (Pazda et al., 2016).

6. *Blue-yellow*

   The last component of the CIE $L^*a^*b^*$ colour space, b^*, was used to measure the amount of yellow in the image. This measure correlates with yellow-blue chroma perception, with higher values of b^* indicating yellow and lower values indicating blue. Both a^* and b^*, therefore, reflect the warmness of colour in the image. For more information on the specifics of colour model conversion, see Ford and Roberts (1998)

7. *Contrast*

   The exact contrast measure used for the purposes of the present experiment was root-mean-square (RMS) contrast. Pixel intensities of all images were firstly normalised in the range [0, 1]. Contrast was then defined as the standard deviation of pixel intensities, where intensities I_{ij} are the i-th and j-th element of an image with size M by N and $\bar{I}$ is the average of all pixel intensities in the image (see the formula below). RMS contrast does not depend on the spatial frequency content of the

image and has been extensively used to investigate the processing of complex stimuli, including natural scenes as well as faces (Kukkonen et al., 1993; Melmoth et al., 2000; Peli, 1990).

$$\sqrt{\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}\left(I_{ij}-\bar{I}\right)^2}$$

8. *JPG size*

   JPG file size has been used as a way of quantifying diversity in image processing and classification as it reflects the amount of information in an image, with larger size being indicative of more diverse content (Deng et al., 2009). This is an artefact of JPG compression which attempts to create patterns the colour values in order to reduce the amount of data that needs to be preserved, thereby reducing the file size. In the context of face perception and evaluation, it is possible that JPG file size reflects skin tone differences where images with a smaller file size present with a more consistent and even skin tone. Looking at the JPG size example in Figure 3.12, we can see that the texture of the image with the smallest JPG file size is much more even. It is also possible that this measure reflects image quality as blurrier images might have less diversity than a sharper high-quality image.

9. *Fourier slope*

   In order to measure the Fourier slope of the images they were firstly resized to 570 x 570 pixels by bicubic interpolation. The log-log frequency spectrum was then determined by computing the discrete Fourier transform where the radially averaged power was plotted as a function of spatial frequency. The slope of this spectrum was then measured by dividing the data points in 30 bins at regular frequency intervals in the log-log plane and performing a least-squares fit of a line to the binned data within the 10-255 cycles/image range. Finally, the slope of the fitted line was calculated. This power spectrum analysis is commonly used for the evaluation of natural scenes (Ruderman &

Bialek, 1994) and artworks (Graham & Field, 2007; Redies, Hänisch, Blickhan, & Denzler, 2007).

10.Blur

The no-reference blur metric proposed by Crete, Dolmiere, Ladret, and Nicolas (2007) was used in order to measure the blur of the images (see Figure 3.15 for a simplified flow chart of the blur estimation principle). This estimation method does not require a reference image as it compares the original version of the image with a blurred version of that same image. To quantify blur we firstly record the intensity variations between neighbouring pixels in the original image, then blur the image using a low-pass filter and record the intensity variations again. The magnitude of the difference between the intensity variations in the original and blurred images can be used as a measure of blur where a bigger difference between the original and the blurred image indicates that the original image was sharper whereas a smaller difference indicates that the original image was already blurred.

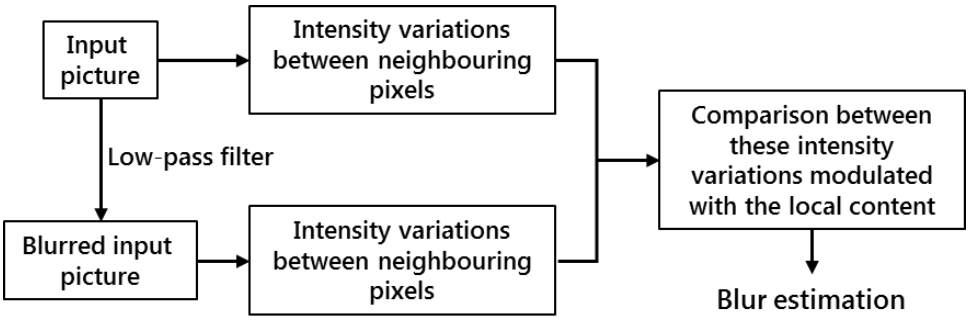

Figure 3.15. Simplified flow of the no-reference blur metric.

11.Face space

Face space was one of the two shape-related measures we used. It refers to the amount of space the face takes up in the whole image. To measure this image/face ratio we used the anatomical landmarks used from Experiment 4. The layout of the landmarks is shown in Figure 3.11 where all numbered points were used to compute the ratio of pixels within and outside the face.

12.Height-to-width ratio

Facial width-to-height ratio is a measure commonly associated with the perception of dominance (Carre & McCormick, 2008). Facial width was measured as the horizontal distance between points 8 and 9 (see Figure 3.11) and facial height was measured as the vertical distance between the midpoint of the horizontal distance between the ridges of valley running from the top lip to the septum (points 25 and 26, Figure 3.16 - bottom), and the midpoint of the horizontal distance between the highest points of the moving part of the lid (points 44 and 55, Figure 3.16 - top).

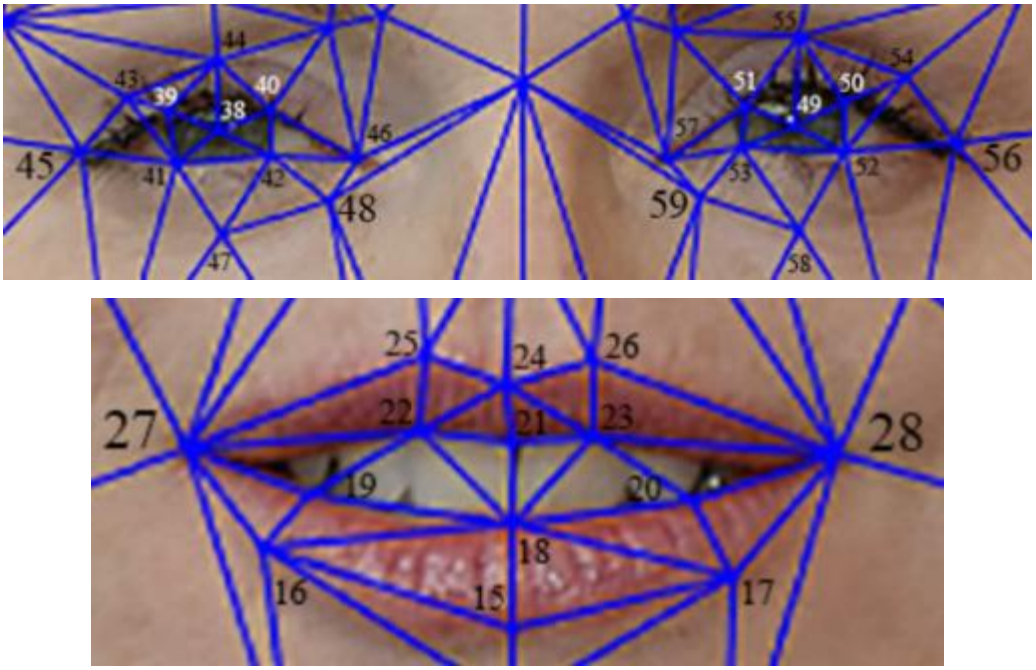

Figure 3.16. Landmark points used to calculate facial width-to-height ratio.

Results and discussion

A standard multiple linear regression was used separately for each social trait (attractiveness, trustworthiness, and dominance) and each identity (100 images per identity). Social attribute judgements were the dependent variable and the physical properties of the images were the

independent variables. These included image saturation, brightness (measured by intensity, value and lightness of colour), contrast, blur, Fourier slope, JPG file size, the amount of redness and yellowness in the face, facial width-to-height ratio and the amount of space the face takes up in the image relative to the size of the image.

Attractiveness

Table 3.4 shows the regression analyses for attractiveness, separately for each of the four identities. The amount of variability in social attribute ratings explained by the physical properties of the images varied greatly across the four identities (R^2_{min} = .13, p > .05; R^2_{max} = .50, p < .001). Common significant predictors of attractiveness judgements included contrast, Fourier slope, JPG file size (all shared between M1 and M2) and facial width-to-height ratio (M2 & F2). Images of these identities were perceived as more attractive with lower contrast, Fourier slope (with lower slope meaning it is closer to -2) and FWHR. The measure of JPG file size is of particular interest as a larger file size is associated with higher attractiveness ratings for one and lower attractiveness ratings for another identity. Such inconsistent results imply that the effect of these physical measures is personal and specific for each identity. This is further supported by the fact that despite these shared predictors, no single measure explained attractiveness ratings significantly for all four identities.

Trustworthiness

Table 3.5 shows the regression analyses for trustworthiness, separately for each of the four identities. Same inconsistent findings were evident for trustworthiness, where the amount of trustworthiness attribution variability explained by the physical attributes of the images varied across identities (R^2_{min} = .16, p > .05; R^2_{max} = .41, p < .001). The only measures that explained trustworthiness judgements significantly and were shared among some identities were contrast (M1 & F1) and Fourier slope (M1 & F2). Images perceived as more trustworthy had lower contrast and slope closer to -2. Again, no single measure explained the variability in trustworthiness ratings

significantly for all four identities, demonstrating that different physical properties of images influence social evaluation to different extends all depending on the specific identity.

Table 3.4. *Summary of Multiple Regression Analyses of Variables Predicting Attractiveness for all Identities.*

	M1	M2	F1	F2
	B (SE)	B (SE)	B (SE)	B (SE)
Saturation	-1.87 (2.28)	6.11 (3.99)	-19.76 (10.96)	-.73 (8.82)
Intensity	-1.35 (4.89)	20.85 (11.38)	21.87 (13.03)	3.37 (11.21)
Value	.50 (3.37)	-10.34 (8.13)	-18.34 (12.00)	-2.97 (7.32)
Lightness	.01 (.10)	-.08 (.11)	-.07 (.25)	.02 (.12)
Red-green	.01 (.04)	.13 (.06)*	.15 (.14)	.04 (.08)
Blue-yellow	.02 (.03)	-.06 (.08)	.43 (.20)*	.04 (.18)
Contrast	-5.23 (2.31)*	-7.64 (3.29)*	-.07 (3.92)	2.69 (1.90)
JPG size	.13 (.05)*	-.06 (.02)**	.01 (.04)	-.02 (.03)
Slope	-2.71 (.99)**	-2.81 (.70)***	-1.22 (1.17)	.23 (1.04)
Blur	.36 (1.59)	-7.96 (1.39)***	-.96 (1.88)	-.25 (1.53)
Face space	-3.29 (1.37)*	-.93 (1.47)	-.15 (1.65)	1.13 (1.30)
W-H ratio	-1.23 (.74)	-3.35 (.93)***	.65 (.83)	-1.73 (.66)*
R^2 (adj. R^2)	.24 (.13)	.50 (.43)	.24 (.13)	.13 (.02)
F	2.27*	7.11***	2.28*	1.12

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 3.5. *Summary of Multiple Regression Analyses of Variables Predicting Trustworthiness for all Identities.*

	M1	M2	F1	F2
	B (SE)	B (SE)	B (SE)	B (SE)
Saturation	-.21 (3.04)	11.70 (5.52)*	-5.38 (11.05)	17.32 (9.34)
Intensity	-4.14 (6.51)	-11.63 (15.77)	8.37 (13.14)	-16.41 (11.87)
Value	-4.19 (4.49)	-6.66 (11.27)	15.92 (12.10)	4.11 (7.75)
Lightness	.14 (.37)	.29 (.15)	-.47 (.26)	.24 (.12)
Red-green	.10 (.05)	.04 (.09)	-.19 (.14)	.03 (.08)
Blue-yellow	.01 (.03)	-.17 (.11)	.04 (.21)	-.37 (.19)
Contrast	-7.01 (3.08)*	.99 (4.56)	-13.79 (3.95)**	-1.26 (2.01)
JPG size	.25 (.06)***	-.01 (.02)	.05 (.04)	-.01 (.03)
Slope	-3.95 (1.31)**	-.60 (.97)	-2.06 (1.18)	-2.62 (1.10)*
Blur	2.32 (2.11)	-1.68 (1.93)	-.15 (1.89)	-3.54 (1.62)*
Face space	-6.34 (1.82)**	2.53 (2.03)	-.99 (1.67)	1.54 (1.37)
W-H ratio	-.39 (.99)	-2.06 (1.28)	.12 (.84)	1.54 (.70)*
R^2 (adj. R^2)	.31 (.22)	.16 (.05)	.26 (.16)	.41 (.33)
F	3.28**	1.42	2.53**	5.06***

* p < .05, ** p < .01, *** p < .001

Dominance

Table 3.6 shows the regression analyses for dominance, separately for each of the four identities. The amount of variability in dominance ratings explained by the physical properties of the images also varied across the four identities (R^2_{min} = .20, p > .05; R^2_{max} = .42, p < .001). Common physical measures that explained dominance ratings significantly included Fourier

slope (M1 & M2) and amount of redness in the face (M2 & F1). Across two out of the four identities, images with a higher amount of redness in the face were perceived as more dominant. The effects of Fourier slope, however, was inconsistent across identities, where a slope closer to -2 was indicative of higher levels of dominance in one identity and lower levels of dominance in another. As with attractiveness and trustworthiness, no single measure was able to predict ratings of dominance significantly for all four identities.

Table 3.6. *Summary of Multiple Regression Analyses of Variables Predicting Dominance for all Identities.*

	M1	M2	F1	F2
	B (SE)	B (SE)	B (SE)	B (SE)
Saturation	.25 (2.40)	-.79 (4.38)	-8.62 (9.18)	-4.69 (9.70)
Intensity	3.71 (5.14)	25.49 (12.52)*	8.24 (10.92)	8.28 (12.32)
Value	2.67 (3.54)	-17.05 (8.95)	-23.90 (10.06)*	-4.48 (8.05)
Lightness	-.10 (.10)	-.08 (.12)	.30 (.21)	-.07 (.13)
Red-green	-.03 (.04)	.15 (.07)*	.23 (.12)*	-.03 (.09)
Blue-yellow	-.01 (.03)	.08 (.09)	.24 (.17)	.15 (.20)
Contrast	4.37 (2.43)	-.55 (3.62)	9.87 (3.29)**	3.36 (2.08)
JPG size	-.13 (.05)*	-.02 (.02)	-.03 (.03)	.01 (.03)
Slope	2.08 (1.04)*	-1.60 (.77)*	.12 (.98)	1.98 (1.15)
Blur	-1.15 (1.67)	-4.64 (1.53)**	-1.14 (1.57)	2.08 (1.68)
Face space	5.71 (1.44)***	-2.90 (1.61)	-.30 (1.39)	-.64 (1.43)
W-H ratio	.66 (.78)	.68 (1.02)	.26 (.70)	-2.74 (.73)***
R^2 (adj. R^2)	.25 (.15)	.26 (.16)	.20 (.09)	.42 (.34)
F	2.43**	2.56**	1.77	5.18***

* $p < .05$, ** $p < .01$, *** $p < .001$

Focusing on the physical measures explaining a significant amount of social attribute variability for each identity, we can see that there are more commonalities within identity, with the same physical measures explaining all three social attributes significantly. For example, face-image ratio, JPG file size and Fourier slope are all significant predictors of attractiveness, trustworthiness, and dominance ratings for M1. What is more, the direction of their relationship is also consistent – images that take up less space in the whole image, with a slope closer to -2 and larger JPG file size were perceived as more attractive and trustworthy but less dominant. Nevertheless, this is not the case for all four identities, demonstrating that the effect of these physical image properties might be tailored to each person specifically. Due to the inconsistencies in the effects of image measures, it is difficult to support any one image property as a reliable predictor of social evaluation.

3.7 General Discussion

In this chapter, we have introduced a data-driven statistical approach which can be used to extract meaningful face variability information and establish the physical correlates of social face attribution. Consistent with previous first impressions studies (Oosterhof & Todorov, 2008; Walker & Vetter, 2009) we demonstrate that there is a high inter-rater consensus in social attribute ratings. This allows us to extract the underlying physical information in the face, diagnostic for social evaluation, and use it to manipulate the perception of the fundamental evaluative dimensions – trustworthiness and dominance, as well as (to a lesser extent) attractiveness. This supports the idea that social face perception is objectively quantifiable. We have also extended previous face evaluation models by incorporating within- and between-person variability together (Experiment 3) as well as exploring the independent effect of idiosyncratic information (Experiment 4) on social evaluation. Our results showed large within-person variability in social attribute ratings which can be used to change the way someone is perceived, just by sampling their own ID-specific variability. In other words, representing someone as having a trustworthy or an untrustworthy face does not tell the whole story. Within each identity, there are images that are perceived as trustworthy- or untrustworthy-looking. Social face evaluation is

therefore not only a function of identity but also a function of the statistical properties of images.

Such data-driven PCA approaches present a number of advantages that make them highly suitable for modelling social face perception. First of all, they do not introduce any theoretical constraints and assumptions but make use of dimensionality reduction techniques that allow the extraction of what is most common in the shape and texture of faces. Furthermore, once face images are subjected to PCA the dimensions of face space, or 'eigenfaces', describe the global properties of these images. Such holistic changes are then applied to the images in order to make them look more or less attractive, trustworthy, or dominant. This is an important point as numerous studies have presented evidence that faces are perceived in a more configural rather than featural way. In the face recognition literature this is illustrated in the classical composite face effect, where the top half of one's face is aligned with the bottom half of another's face which is shown to interfere with the recognition of both identities (Young et al., 1987). The effect has been replicated in various contexts including the perception of emotional expressions (Calder, Young, Keane, & Dean, 2000), face gender (Baudouin & Humphreys, 2006), and face race (Michel, Corneille, & Rossion, 2007). More recently, Todorov et al. (2010) utilised an adapted version of the composite face effect where participants were presented with composite faces made from trustworthy and untrustworthy halves and asked to evaluate either the top or the bottom half of the face while ignoring the other one. Regardless of the instructions they received, participants rated the same halves more highly when they were aligned with trustworthy compared to untrustworthy complements, which implies that social attribute judgements also rely on holistic processing.

The approach we have employed here adds to existing face evaluation models. We utilised natural, highly variable images which are more representative of everyday encounters than the stimuli typically used in psychological research. Use of these 'ambient images' provide valuable data in that they allow statistical extraction of dimensions underlying natural

variability. The novel aspect of our approach is the integration of within-person variability by sampling different images of the same identity. A number of studies have already demonstrated that within-person variability in certain social attribute ratings exceeds or at least is comparable to between-person variability (Jenkins et al., 2011; Todorov & Porter, 2014) and here we demonstrate that this ID-specific within-person variability is sufficient to allow the manipulation of the way people are perceived. This supports the idea that we need a full account of both within- and between-person variability in order to understand face perception completely.

A particularly striking finding replicated in both Experiments 3 and 4 is the difference in social evaluation for the two fundamental dimensions identified by Oosterhof and Todorov (2008), compared to evaluations of attractiveness. This is in contrast to Sutherland et al. (2013) who extracted an additional factor – youthful-attractiveness using naturally occurring 'ambient' face images. However, as the name suggests, it is possible that this factor relies on age differences to a greater extent. Indeed, recent studies have shown a clear distinction in attractiveness judgements compared to valence and dominance, suggesting attractiveness is evaluated in a rather distinct way. Todorov and Porter (2014), for example, report attractiveness as the only trait with between-person variability exceeding within-person variability and Sutherland, Young, and Rhodes (2017) also show significantly less variance in ratings of attractiveness, especially for male identities. This interpretation is further supported by data from the original face evaluation model by Oosterhof and Todorov (2008). While many of the traits with a positive loading on one dimension have a negative loading on the other dimension, attractiveness and confidence present with relatively high positive loadings on both dimensions, implying some underlying mechanism differences. Finally, Sutherland et al. (2015) used ratings of youthful-attractiveness, approachability and dominance to predict traits from the Big Five model (extraversion, agreeableness, openness, neuroticism and conscientiousness; McCrae & Costa, 1987). They showed that four out of the five traits were best explained by the approachability factor and dominance was the best predictor of conscientiousness. Youthful-attractiveness, on the other hand, was not a

good predictor of any of the Big Five traits, further highlighting the distinction between attractiveness and other social evaluation traits. Together with the present findings that perceptions of attractiveness were particularly difficult to manipulate using idiosyncratic variability only, these studies suggest that attractiveness depends on identity to a large extent, and that it is evaluated in a qualitatively different way than other social attributes.

The exploratory analysis of low-level image properties and their effect on social evaluation described in Experiment 7 presents a very inconsistent pattern of results across identities and social attributes. Results within each identity provided some support for the influence of certain image properties on first impressions. Facial contrast and a Fourier slope closer to that of aesthetically pleasing images (-2), for example, were significant predictors of attractiveness supporting findings from Menzel et al. (2016) and Jones et al. (2015). Despite the great many studies demonstrating the effect of facial width-to-height ratio on dominance perception (Carre & McCormick, 2008), it did not predict dominance attribution consistently in Experiment 7. This is possibly due to the large differences in the ratio across images of the same identity as well as the use of faces displaying different emotional expressions, which can disrupt the FWHR greatly. Overall, there was not a single measure that predicted any of the fundamental social dimensions consistently and certain physical properties influenced attribution of all three social dimensions within a single identity (e.g., contrast and slope predicted all social traits for M1 and FWHR predicted all traits for F2). This implies that the effect of such low-level properties might be more idiosyncratic than previously thought.

Another interesting aspect of social face attribution that can be addressed and extended with the present design is the face overgeneralisation effect. This refers to people's tendency to interpret any transient behavioural changes such as emotional expressions as stable enduring personality attributes (Secord, 1959). Montepare and Dobish (2003), for example, demonstrated that neutral faces that resemble a happy expression elicit impressions of high affiliative traits (comparable to impressions of

trustworthiness in face evaluation literature) while faces resembling an angry expression elicit impressions of high dominance. These findings are further supported by Oosterhof and Todorov (2008) who showed that manipulating neutral computer-generated faces to look more trustworthy made them look happier and manipulating these same faces to look less trustworthy made them look angrier. Using a Bayesian network expression classifier, Said et al. (2009) also found a high positive correlation between positive face evaluation attributes and the probability of classifying faces as expressing happiness, as well as a high correlation between the probability of classifying faces as angry and judgements of dominance and aggressiveness. Looking at the reconstructed pairs of images manipulated to look more and less attractive, trustworthy, and dominant, it is clear that emotional overgeneralisation effects play an important role in social face evaluation within each identity, especially for trustworthiness and dominance judgements. Taken together, evidence from both neutral and expressive faces have implications for dual models differentiating between processes involved in the perception of stable face properties such as identity and processes involved in the perception of changeable face properties such as emotional expressions (Bruce & Young, 1986; Haxby, Hoffman, & Gobbini, 2000). Findings that people make use of transient states to infer stable personality characteristics, however, imply that such processes are not fully independent and interact with one another.

In conclusion, we have demonstrated the importance and magnitude of within-person variability in social face evaluation. Different images of the same identity gave rise to very different social attribute judgements and this idiosyncratic information was sufficient to capture the underlying physical information people use to inform their judgements as well as bring about significant changes in social evaluation with no changes in identity. Moreover, we show that within-person variability can be extracted for each identity, and used to reconstruct both the original and novel images and change how they are evaluated on socially-important dimensions. This chapter describes the first attempt at incorporating ID-specific and shared variability in order to manipulate social evaluation, addressing the idea that the attribution of socially important traits such as trustworthiness and

dominance depend on both identity and the specific images used. This suggests that social evaluation is not only a function of identity but also a function of the statistical properties of face images.

Chapter 4 – First Impressions in Face Matching

4.1 Introduction

The faces we see in our everyday lives allow us to extract both identity- and emotion-related information, such as age, gender, mood, and even personality with varying levels of accuracy and agreement (Albright et al., 1997; Bruce & Young, 1986; Rule et al., 2013; Todorov et al., 2015). Fundamental face perception models, however, argue that identity, emotion, and speech reading are processed somewhat independently of one another when processing familiar faces (Bruce & Young, 1986). This functional independence is supported by both behavioural and neuropsychological findings. For example, studies have reported a familiarity advantage for identity-, but not expression-matching tasks (Bruce, 1986; Young, McWeeny, Hay, & Ellis, 1986) and there are reports of brain-lesioned patients exhibiting relatively selective impairments in identity, emotion, and speech processing (Humphreys, Donnelly, & Riddoch, 1993; Parry, Young, Shona, Saul, & Moss, 1991; Young, Newcombe, de Haan, & Hay, 1993). Unfamiliar faces, however, cannot be associated with the same identity-specific semantic codes (e.g. name, occupation) as familiar faces and rely mostly on purely pictorial or visually-derived semantics (Hancock, Bruce, & Burton, 2000; Megreya & Burton, 2006). Therefore, they might be affected by changes in expression, pose and other pictorial factors to a greater extent.

Unfamiliar faces are also more commonly linked to social evaluation with people shown to attribute social characteristics such as trustworthiness and dominance to unfamiliar faces automatically and within a few milliseconds (Willis & Todorov, 2006). Research on such first impressions has demonstrated that, just as unfamiliar face recognition, social attribution depends on pictorial factors such as emotional expressions, eye gaze, and image contrast (Bayliss & Tipper, 2006; Russell, 2003; Said et al., 2009). Guided by the automatic nature of social evaluation and the similarities in the factors affecting both unfamiliar recognition and attribution, the experiments in this chapter aim to investigate the relationship between

recognition performance and the main dimensions of social evaluation – trustworthiness, dominance, and attractiveness (Oosterhof & Todorov, 2008; Sutherland et al., 2013).

Face recognition tests

There are two main approaches to testing face recognition that tap into different cognitive processes. Originally based on classical word and object recognition models (Nelson, Reed, & McEvoy, 1977; Warren & Morton, 1982), face memory is tested through old/new paradigms where participants are instructed to learn a number of faces and are later presented with the same 'old' faces mixed in with some 'new' faces. Their task then is to indicate whether they have seen those faces during the learning phrase or not. In contrast to face memory tasks, face matching is a mainly perceptual task. Here, participants are presented with a pair of face images and asked to decide whether they are of the same identity or of two different identities (Clutterbuck & Johnston, 2002, 2004). Trials where participants are presented with two different images of the same person are referred to as match trials, whereas images of two different identities are used in mismatch trials. The most commonly used memory test is the Cambridge Face Memory Test (CFMT, Duchaine & Nakayama, 2006), and the Glasgow Face Matching Test (Burton et al., 2010) is a standard validated face matching test based on the earlier applied work of Bruce et al. using line-up tasks (1999). With recognition paradigms being developed earlier and memory processes receiving more research attention in the past, a lot more is known about the potential factors influencing face recognition. Gaining the same amount of understanding about face matching will therefore allow us to explore identity recognition further and to distinguish factors that influence face memory from those associated with perceptual processes only.

An interesting characteristic of unfamiliar face matching and memory tasks is that they do not give rise to a mirror effect (Glanzer & Adams, 1985, 1990). This is surprising as this regularity applies to the recognition of many other stimulus categories such as words and everyday objects, where items accurately recognised as old when used as targets are also accurately rejected

as new when used as distractors (see Glanzer, Adams, Iverson, & Kim, 1993 for a review ; Snodgrass & Corwin, 1988). Faces, on the other hand, present no association between hits and false positives (FPs) when used in a recognition task (Bruce, Burton, & Dench, 1994; Hancock et al., 1996) and no correlation between match and mismatch trials when used in a matching task (Megreya & Burton, 2007). As familiarity is one of the key factors in both face memory and face matching (Bruce et al, 1999, 2001; Megreya & Burton, 2006, 2008), it is important to note that this is true for unfamiliar faces only. Studies using familiar or even familiarised faces have reported a large significant correlation between hits and FPs, implying that unfamiliar faces are processed in a qualitatively different way than familiar faces (Megreya & Burton, 2006, 2007).

Factors affecting face matching

In addition to familiarity, image variability has also been shown to be an important factor in face matching. It incorporates long- and short-term changes in the person (e.g. aging and emotional expressions), changes in the world (e.g. lighting and camera angle), as well as changes in capture devices (e.g. resolution and focal distance). Starting with differences in the person, Megreya, Sandford and Burton (2013) used a 1-in-10 face matching line-up task where different images of the same person were either taken on the same day or an average of 17.2 months later. Results indicated a much lower matching accuracy for images taken further apart in time, demonstrating how a relatively short amount of time can produce image variability that affects face matching performance. In fact, studies report a decrease in matching accuracy with even momentary changes such as emotional expressions (Chen, Lander, & Liu, 2011). Bruce, for example, showed that a mismatch in emotional expression between images of the same person could impair performance in both face recognition (Bruce, 1982) and 1-in-10 matching tasks (Bruce et al., 1999). Hill and Bruce (1996) extended these findings further by incorporating changes in the person and the outside world together. They explored face matching across changes in viewpoint (from a ¾ view to a profile and vice versa) and lighting (either from above or below) demonstrating near ceiling accuracy for images taken under the same

conditions and a significant decrease in performance when viewpoint or lighting was manipulated. Finally, matching high- and poor-quality images of unfamiliar identities can also impair recognition with 90% accuracy for matching trials with two high-quality images compared to 70% for trials where one of the images was pixelated (Bindemann, Attard, Leach, & Johnston, 2013; Burton et al., 1999).

Methods for improving unfamiliar matching performance

Evidence of the frailty of face matching performance has motivated empirical work to establish ways of improving recognition accuracy using a range of techniques. This is important as establishing successful training procedures could potentially bring about significant improvements in person identification as well as national security. Unfortunately, evidence for the benefits of training is limited, at best, for both intensive (Woodhead, Baddeley, & Simmonds, 1979) and short-term (Dolzycka, Herzmann, Sommer, & Wilhelm, 2014) training courses. Focusing on face shape specifically and adopting a feature-by-feature comparison approach has been shown to improve face matching accuracy in both student and expert facial examiner samples (Towler et al., 2017). This improvement, however, was seen in match trials only and other shape-related strategies such as classifying the shape of the head as square, oval or round did not produce any significant improvements in performance (Towler, White, & Kemp, 2014).

A different approach to improving face recognition accuracy is providing trial-by-trial feedback as it could alert participants to the unexpected difficulty of the task. White, Kemp, Jenkins, and Burton (2014b), for example, incorporated trial-by-trial feedback in the Glasgow Face Matching Test and showed cumulative improvements in accuracy. However, some argue that feedback does not improve performance *per se* but rather prevents participants from making more mismatch errors (Alenezi & Bindemann, 2013). Another strategy that has been shown to improve certain aspects of face recognition is incorporating within person variability. This is achieved by presenting multiple images of the target identity, which helps participants gather vital information about the way this person may vary.

While feedback was shown to help mismatch trials only, presenting participants with two different images of the same person has been shown to improve performance in match trials only (White et al., 2014a). Moreover, taking the variability approach a step further by creating averages of many images of the same person preserves any identity-specific information, while removing image-specific information (Burton et al., 2005). Using such an average image has been shown to improve matching performance, however, it is not clear whether averaging produces any additional improvement in accuracy over that achieved by using multiple images per identity (Burton et al., 2011). Finally, a simple manipulation that has been shown effective in face matching tasks is pairing participants together and asking them to come to a joint decision (Dowsett & Burton, 2015). Here participants performed three face matching tests – a test measuring performance as a pair as well as an individual face matching test before and after pairing that aimed to investigate any carryover effects. Identifying the high and low performers in each pair then showed a significant improvement in the individual matching accuracy of low-scoring participants after completing the task together with their counterparts, compared to their initial individual performance.

Overview of studies

As with person recognition, social face evaluation has been shown to depend on identity (Oosterhof & Todorov, 2008), image properties (Jenkins et al., 2011; Todorov & Porter, 2014) and especially emotional expressions (Said et al., 2009). It is therefore possible that trait attribution can affect identity recognition to a certain extent. Support for this suggestion comes from studies investigating face memory. Some, for example, report higher recognition accuracy when participants were asked to rate faces on social traits such as intelligence and likeability rather than simply evaluate the physical properties of the face (Bower & Karlin, 1974; Courtois & Mueller, 1979; Winograd, 1976). The association between social attribution and face matching performance has not been explored yet and a different pattern of results might be expected for match and mismatch trials. On one hand, as both images in the face pair are of the same identity in match trials, it is expected that smaller differences in social attribute ratings might lead to a

more accurate performance on those trials. On the other hand, images in mismatch trials are of two different identities, therefore, a greater difference in social attribute ratings might be expected to improve face matching accuracy. Experiment 8 investigates these predictions by collecting social ratings for all images used in a matching task and correlating the social rating differences with matching performance. Experiment 9 takes the idea further and introduces a specific context by embedding images in a passport frame. Finally, Experiments 10 and 11 focus on the effect of emotional expressions on face matching and the information provided through a smile, in particular, as a possible way of improving face matching accuracy.

4.2 Experiment 8

Introduction

With research demonstrating that faces are socially evaluated automatically and within a few milliseconds (Willis & Todorov, 2006), it is possible that people might use these social attributions as cues to identity recognition. There is already some evidence for this coming from face memory studies where rating faces on social attributes in the learning phase led to higher accuracy in the recognition phase. This was contrasted with gender classification or rating single features of the face during learning, neither of which improved recognition accuracy (Winograd, 1976). The association between identity and social attributes has not been explored yet using matching tasks which, compared to face recognition, are perceptual and do not depend on memory. Such an investigation will therefore allow us to establish whether social evaluation improves face memory only or can also be related to our perceptual memory-independent identity decisions.

Influential face evaluation models identify two fundamental dimensions of social attribution – trustworthiness and dominance (Oosterhof & Todorov, 2008). Further studies using naturally-occurring images replicate this structure and identify an additional dimension, referred to as youthful-attractiveness (Sutherland et al., 2013). Besides the type of stimuli it is based

on, attractiveness evaluation seems to be qualitatively different from trustworthiness and dominance attribution. Todorov and Porter (2014), for example, compare the variability of social ratings within (different images of the same person) and between (images of different people) identity. They report a distinct pattern of results for attractiveness judgements which compared to all other social attributes presented with greater between-, rather than within-, person differences. This, together with findings from Chapter 3, implies that ratings of trustworthiness and dominance rely on both identity and image properties whereas attractiveness judgements are more consistent within identities.

This experiment aims to explore the relationship between social evaluation and face matching performance. Based on existing face evaluation models, images were rated for trustworthiness, dominance, and attractiveness (Oosterhof & Todorov, 2008; Sutherland et al., 2013). These same images were then used in a matching task with the hypothesis that social ratings will be related to matching performance. There were two distinct predictions for match and mismatch trials. As participants see two different images of the same person during match trials, images with more similar social ratings will produce more 'same' responses which will improve overall matching performance. During mismatch trials, on the other hand, participants are presented with images of the two different identities. Here, images with similar ratings that trigger more 'same' responses will lead to more errors. Therefore, a larger difference between the social ratings attributed to images used in a match trial will decrease performance, whereas a larger difference between the social ratings attributed to images used in a mismatch trial will improve face matching performance.

Attractiveness, trustworthiness, and dominance were the traits included in this analysis as they have been identified as the fundamental dimensions of social evaluation. We had no distinct predictions about each social trait, however, following from the results reported in Chapter 3 as well as previous findings about the dissociation between trustworthiness and dominance on one hand, and attractiveness, on the other, it is possible that

they are associated with face matching to a different extent (Todorov & Porter, 2014). Findings from Experiment 3 showed a lot more between-person variability in ratings of attractiveness, compared to ratings of trustworthiness and dominance. This implies that attractiveness might be a more salient cue to identity and therefore could be of more importance when it comes to identity tasks such as face matching.

Method

Participants

A total of 80 participants (20 male, mean age = 20.4, age range = 18-28) completed the face matching task and a different sample of 38 participants (6 male, mean age = 20.7, age range = 18-26) completed the image rating task. All had normal or corrected-to-normal vision and received payment or course credits for their participation. Informed consent was provided prior to participation and experimental procedures were approved by the ethics committee of the Psychology Department at the University of York.

Materials

A total of 240 face images were used as experimental stimuli. These comprised four different images of 40 unfamiliar identities (20 male, same image set used in Experiment 1). In order to use these stimuli in a matching task, one extra match and one foil image were collected for each identity. Match images were gathered from Google Image Search on the names of target identities. Foil images were of a different identity that resembled the target identity closely and matched its verbal description. All images were in full colour, broadly frontal-view, and with no parts of the face obscured by clothing or glasses. They were all 'ambient', naturally occurring images and captured a good amount of face variability due to differences in lighting, pose, and emotional expressions (Jenkins et al., 2011). Face images were cropped so no background information was available and resized to 190 x 285 pixels. See Figure 4.1 for examples.

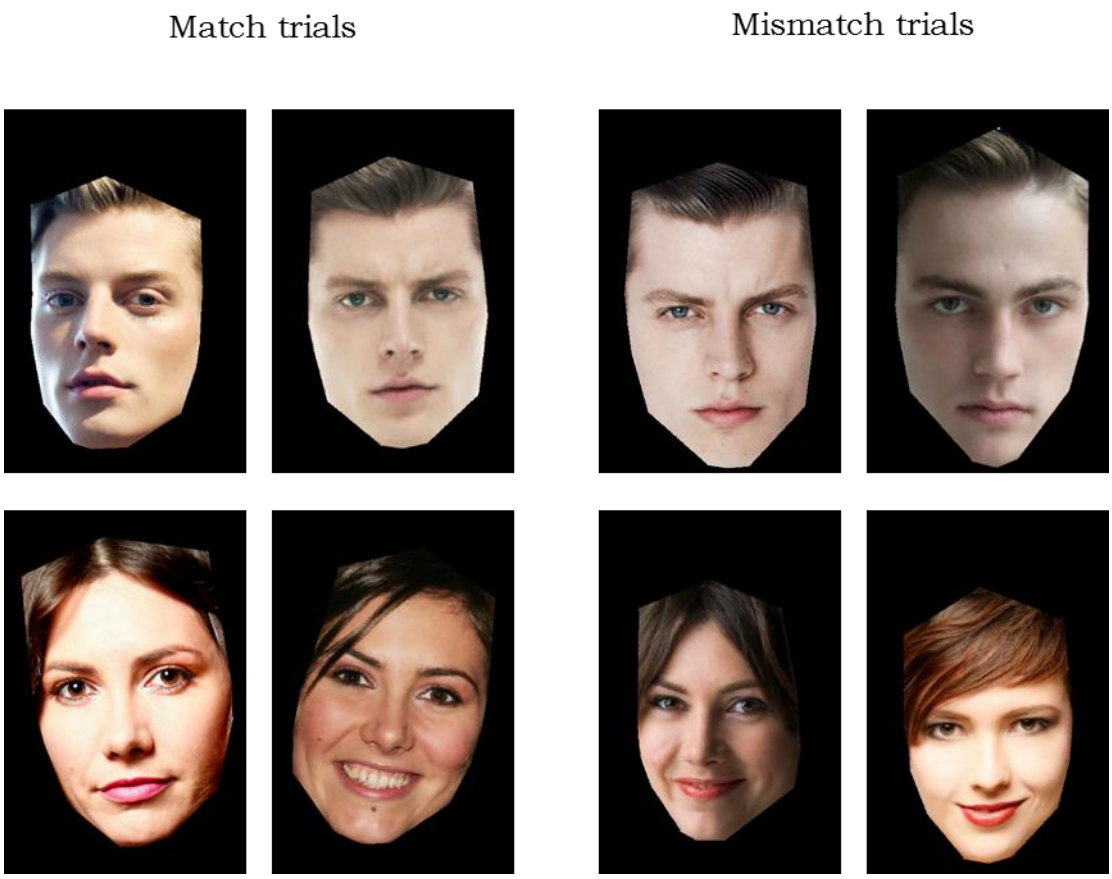

Figure 4.1. Examples of the experimental stimuli and trial structure. On match trials images are of the same identity and on mismatch trials images are of two different identities. Here, the first three images in each row are of the same identity, followed by the foil image.

Procedure

The study took place in a room equipped with a standard PC running MATLAB R2014a. Stimuli were displayed on an 18-inch monitor and the experimental program was written in MATLAB using functions from the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). For the matching task, participants completed 80 trials with an equal number of match and mismatch trials. Participants were presented with two images on the screen and asked to decide whether these images were of the same identity or two different identities by pressing the corresponding keys on the keyboard. On each trial participants saw one of the four exemplars of a particular identity alongside the match (same identity) or mismatch (different

identity) image (see Figure 4.1 for an example). The task was not timed but participants were instructed to be as quick and accurate as possible. Each participant saw each identity twice – once in a match and once in a mismatch trial. However, exemplars were counterbalanced so that participants never saw the same image twice. Trial order was randomised independently for each participant.

For the rating task, participants were asked to rate each image for attractiveness, trustworthiness, and dominance on a scale from 1 (not at all) to 9 (extremely). Each face was presented at the centre of the screen with the rating scale positioned below each image. Ratings for each trait were collected in separate blocks to avoid any carryover effects (Rhodes, 2006). Participants were also encouraged to rely of their first impressions (or 'gut feeling') rather than spend much time evaluating each image. Block and image presentation order was randomised.

Results and discussion

Differences in attribute ratings for match and foil images

Ratings of all social attributes showed good inter-rater reliability with all Cronbach's alphas above .80 (Nunnally, 1978), so ratings for each image were averaged across participants separately for each trait. Figure 4.2 shows rating differences between face pairs in match and mismatch trials. A 3x2 within subjects ANOVA (attribute: attractiveness vs trustworthiness vs dominance; trial type: match vs mismatch) showed a significant main effect of social trait (F (2, 318) = 8.59, p < .001, η_p^2 = .05) and a significant main effect of trial type (F (1, 159) = 23.28, p < .001, η_p^2 = .13). These main effects were qualified by a significant interaction (F (2, 318) = 11.50, p < .001, η_p^2 = .07). Simple main effects revealed that image rating differences in match trials were significantly smaller than image rating differences in mismatch trials for attractiveness evaluation (F (1, 477) = 45.75, p < .001, η_p^2 = .09). Rating differences in match and mismatch trials were not significantly different for trustworthiness (F (1, 477) = 1.79, p < .05, η_p^2 < .01) or dominance evaluation (F (1, 477) = 1.13, p < .05, η_p^2 < .01).

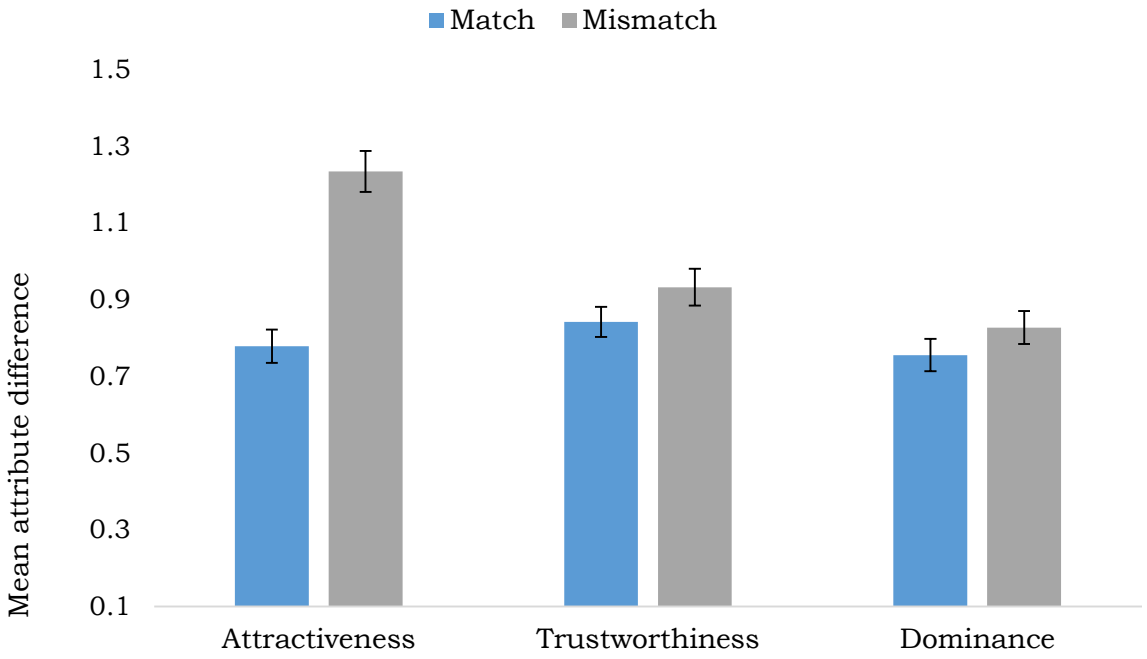

Figure 4.2. Mean attribute difference between match and mismatch pairs. Error bars represent within-subjects standard error (Cousineau, 2005).

Results demonstrate that ratings of attractiveness can reflect differences in identity, as there were larger discrepancies in attractiveness ratings between images in mismatch, rather than match trials. This is consistent with findings from Chapter 3 demonstrating a lot more between-person variability in ratings of attractiveness than ratings of trustworthiness and dominance and making attractiveness a much more reliable cue to identity.

Correlations between matching accuracy and attribute differences

In order to explore the relationship between social attribute ratings and face matching accuracy, the difference in ratings between the images in each face pair (attribute difference = exemplar rating - match/foil rating) was correlated with matching accuracy. As discussed above, we predict that similar ratings will lead to good performance for matching pairs, but poor performance for mismatching pairs because similarities in social ratings will

lead to more 'same' responses. As there was a different set of predictions for match and mismatch trials data were analysed separately. Figure 4.3 uses data from the experiment to illustrate each prediction where the mean attractiveness rating is shown above each image. For match trials images with similar attractiveness ratings (top left) were accurately identified as images of the same person 95% of the time, whereas images with different attractiveness ratings (bottom left) were identified as images of the same person only 40% of the time. The inverse is true for mismatch trials – here, images with different attractiveness ratings (bottom right) were accurately identified as being of two different identities 95% of the time, whereas images with similar attractiveness ratings (top right) were identified as being two different identities only 30% of the time.

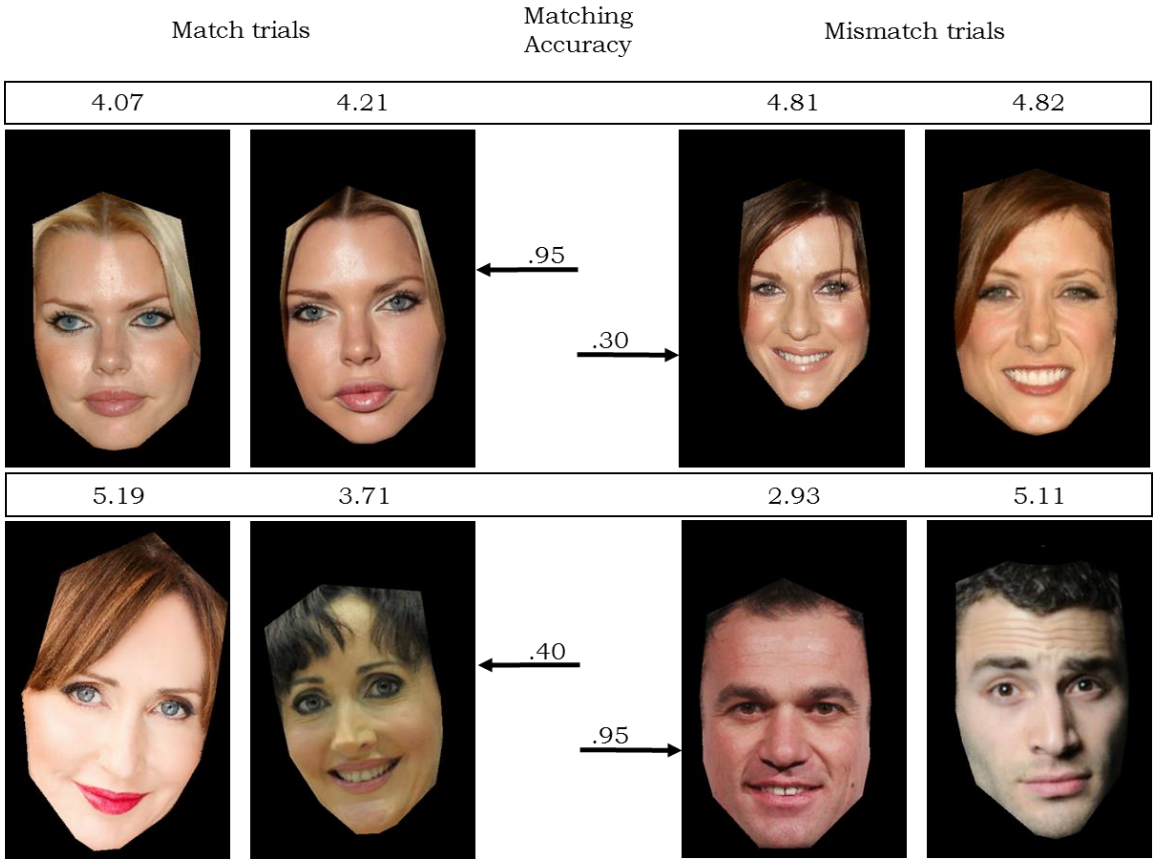

Figure 4.3. Example of predictions for match (left) and mismatch (right) trials. The graph shows attractiveness ratings and matching accuracy for each particular trial.

Tables 4.1 and 4.2 show the correlations between face matching accuracy and differences in social ratings for match and mismatch trials respectively. Pearson's correlations revealed no significant relationships between any attribute difference and face matching performance in match trials, however there was a significant positive correlation between attractiveness difference and matching performance in mismatch trials, r (160) = .17, p < .05. The direction of this relationship fits our prediction as it shows that the bigger the difference in attractiveness rating between images, the more likely it is that people will accurately determine that these are actually images of two different identities. Moreover, there was a significant positive correlation between attractiveness and dominance differences in both match, r (160) = .23, p < .01 and mismatch, r (160) = .18, p < .05 trials.

Table 4.1. *Correlations Between Face Matching Accuracy and Differences in Social Attribute Ratings in Match Trials.*

	Attractiveness Difference	Trustworthiness Difference	Dominance Difference
Accuracy	-.13	.08	-.11
Attractiveness Difference		-.01	.23**
Trustworthiness Difference			-.11

Table 4.2. *Correlations Between Face Matching Accuracy and Differences in Social Attribute Ratings in Mismatch Trials.*

	Attractiveness Difference	Trustworthiness Difference	Dominance Difference
Accuracy	**.17***	-.05	.11
Attractiveness Difference		.02	.18*
Trustworthiness Difference			.08

Overall, results did not show much support for the association between social evaluation and matching performance. With the exception of attractiveness, differences in social traits were not used as cues when matching unfamiliar faces. Despite some evidence that participants can detect differences in identity with judgements in attractiveness, they did not seem to use this information when matching images of the same identity. The only association between social evaluation and matching accuracy was found for ratings of attractiveness, specifically on mismatch trials. This implies that differences in social attributes are only relevant in situations that encourage participants to look for differences between identities rather than similarities. It is therefore possible that after an identity decision has been made, participants use attractiveness information in order to support or discard their initial response. Relating these findings to previous literature demonstrates clear differences in the relationship between social evaluation and face *recognition*, on one hand, and social evaluation and face *matching*, on the other. While social judgements bring about advantageous depth of processing when learning and remembering faces (Bower & Karlin, 1974; Courtois & Mueller, 1979), the link between social attribution and perceptual matching is only minimal.

4.3 Experiment 9

Introduction

The next experiment aimed to replicate these findings and explore any potential effects of context, which could prime the importance of specific social traits. The influence of context on face recognition has been demonstrated in both memory recognition and perceptual matching tasks. Rainis (2001), for example, showed an improvement in recognition accuracy when faces were seen in the same context at both the learning and recognition phases, and later studies replicated these findings even when attention was not directed towards the context (Shriver, Young, Hugenberg, Bernstein, & Lanter, 2008).

The mere addition of frame can affect recognition, as demonstrated by studies embedding face images in newspaper articles and manipulating the valence of the newspaper headlines (Galli, Feurra, & Viggiano, 2006). Findings from this study showed a memory advantage for faces learned in a more negative context and more importantly, that faces embedded in a newspaper frame were generally better remembered than faces learned in isolation. Not only does context influence face recognition memory, but its impact is also present in perceptual matching tasks. A recent study by McCaffery and Burton (2016) examined face matching when one of the images was presented within a passport frame in order to simulate a real-life identity check situation. Results showed a systematic bias such that participants were poorer at detecting a mismatch when one of the faces was embedded in a passport frame. It is possible that the passport frame gives images a more official and legal quality, making them difficult to discard as invalid or fraudulent.

Context is also an important factor in social evaluation, with studies showing different sets of social traits being important for different social contexts and situations. Todorov and Porter (2014), for example, showed participants multiple images of the same identity and asked them to select the most suitable image for a number of different contexts, including a job application, an online dating website, and a political campaign poster. A different pattern of social traits was critical for different scenarios – the strongest predictors for job applications were trustworthiness, competence, and intelligence whereas trustworthiness, extraversion and meanness (negatively correlated) were better predictors of images chosen for dating websites.

Incorporating evidence for the effect of context on both face recognition and social evaluation, this study aimed to replicate findings from Experiment 8 in a more applied context. Here the match and foil images were embedded in a passport frame to simulate a passport check situation, which could potentially activate trustworthiness, a judgement relevant to checking ID in a formal context.

Method

Participants

A total of 80 participants (11 male, mean age = 19.7, age range = 18-27) from the University of York took part in the study. All had normal or corrected-to-normal vision and received payment or course credits for their participation. Informed consent was provided prior to participation and experimental procedures were approved by the ethics committee of the Psychology Department at the University of York. Only participants who had not taken part in Experiment 8 were recruited for the present experiment.

Materials

The same images were used as in Experiment 8 with the only difference being that match and foil images were embedded in an American passport frame. Exemplar images were resized to 195 x 262 pixels and passport frames were resized to 620 x 429 pixels. Both face images in each pair were the same size. Biographical information, place of birth, date of issue, date of expiration, given name, and surname were assigned as follows. Dates of issue ranged from 2007 to 2014 and dates of expiration ranged from 2017 to 2024. Given names and surnames were randomly chosen from the most common American names. As these identities are celebrities in other countries (unfamiliar to UK participants) it is possible to access their own dates of birth so these were used in the passport frames. The same passport information was used for match and foil images of the same identity. Passport frame images were created with Corel Paintshop Pro X6 using a blank American passport frame. An example can be seen in Figure 4.4.

Procedure

The experiment followed the same procedure as the matching task in Experiment 8. Participants were instructed not to take the data in the passport into consideration when completing the matching task.

Figure 4.4. Example of a match and mismatch trial for the same identity.

Results and discussion

In order to explore the relationship between social evaluation and face matching accuracy the absolute attribute difference for each match and mismatch pair was firstly calculated across all identities and three social traits. Tables 4.3 and 4.4 show Pearson's correlations between these differences and matching accuracy in match and mismatch trials respectively. Again, as we have distinct predictions for both types of trials, they were analysed independently.

Table 4.3. *Correlations Between Face Matching Accuracy and Differences in Social Attribute Ratings in Match Trials.*

	Attractiveness Difference	Trustworthiness Difference	Dominance Difference
Accuracy	-.08	.09	-.08
Attractiveness Difference		-.01	.23**
Trustworthiness Difference			-.11

Table 4.4. *Correlations Between Face Matching Accuracy and Differences in Social Attribute Ratings in Mismatch Trials.*

	Attractiveness Difference	Trustworthiness Difference	Dominance Difference
Accuracy	**.18***	-.11	**.16***
Attractiveness Difference		.02	.18*
Trustworthiness Difference			.08

Consistent with findings from Experiment 8, no significant correlations were found between attribute differences and matching performance in match trials. A different pattern of results was found for mismatch trials with significant positive correlations between matching performance and both attractiveness (r (160) = .18, p < .05) and dominance differences (r (160) = .16, p < .05). Also replicating results from Experiment 8, there was a significant positive correlation between attractiveness and dominance differences for both match (r (160) = .23, p < .01) and mismatch trials (r (160) = .18, p < .05).

The introduction of context was intended to present participants with a different situation where the importance of social traits other than attractiveness might be primed as a relevant cue for identity decisions. The passport control context was specifically chosen to activate the importance of trustworthiness. Embedding match and foil images in passport frames replicated findings from Experiment 8 that highlight the importance of attractiveness for mismatch trials. Moreover, greater differences in dominance ratings of images used in mismatch trials were also associated with more accurate face matching performance. This is surprising given our prediction about the salience of trustworthiness in this context. When general social perception models are related to threat evaluation, trustworthiness is described as someone's intent to do harm, whereas dominance is described as their ability to act on such intentions (Fiske et al., 2007; Oosterhof & Todorov, 2008). It is, therefore, possible that dominance is the more critical trait in this context as it could lead to worse consequences. Nevertheless, while this additional relationship might have been identified with the introduction of a more applied context, any interpretation should be treated with caution due to the small correlation coefficients. Differences in social attribute ratings seem to be relevant to performance on mismatch trials only, which to a certain extent supports the role of context. Once participants have made an initial decision that the images presented are of two different identities, they are more likely to look for differences, including those in social traits, in order to justify their decision. If participants decide images depict the same identity, however, they might be more likely to look for similarities instead.

4.4 Experiment 10

Introduction

Emotional expression is a factor related to both recognition and social evaluation, although its effect on the latter is much stronger as demonstrated by studies on emotion overgeneralisation (Zebrowitz et al., 2010). Said et al. (2009), for example, showed that faces with a subtle resemblance to happy

faces are perceived as more sociable, trustworthy, and caring, whereas faces similar to angry faces are seen as more dominant and aggressive. The effect of emotional expression is much more subtle in face recognition. Bruce and colleagues (1999) used faces with different emotional expressions in a matching task and reported a decrease in accuracy for emotion incongruent pairs where one image displayed a happy and the other a neutral expression.

Despite the results from Experiments 8 and 9 showing little evidence for the association between first impressions and face matching, it is still possible that some key aspects of social evaluation will be relevant to matching performance. As we express different emotions our faces reveal information that reflects both anatomical changes in the positioning of bones or muscle contractions as well as idiosyncratic activation patterns related to specific emotions. A smile is one of the most common and universally recognised emotional expressions, especially in the context of first impressions. Smiling has been shown to involve two facial muscles – zygomaticus major whose contraction pulls lip corners up and orbicularis oculi whose contraction leads to changes in the eye region such as wrinkles in the eye corners (crow's feet), narrowing of the eye opening, and bags under the eyes becoming more pronounced (Ekman, 1992). While most identity recognition research has been focused on faces in their neutral state and a lack of emotional expression is required when using face images in an official capacity (e.g. in passports and national IDs), it is possible that expressive and smiling faces in particular might reveal further underlying diagnostic information about individuals making them easier to recognise.

Evidence for this suggestion comes from the automatic face recognition literature where different computational algorithms are used to maximise recognition accuracy. Yacoob and Davis (2002), for example, used a PCA-based algorithm with neutral, angry, and happy faces and demonstrated that expressive faces had higher discrimination power, meaning that identities were recognised to greater extent when an expressive image was used to represent them in the algorithm. Moreover, using the PCA components to reconstruct the angry and happy faces in the 'neutral face space' and vice

versa showed that expressive faces had higher discrimination power when projected in the neutral space compared to neutral faces projected in the expressive space. Such findings imply that expressive faces provide some extra identity-diagnostic information that can enhance recognition, at least computationally. This is further supported by meta-analytic studies that explore the key factors affecting recognition performance by comparing different face recognition algorithms. A consistent finding is that recognition is significantly impaired when the target and query images express different emotions (Lui et al., 2009). In emotion-congruent sets, however, all algorithms have higher estimated probability of verification when the target and query faces are smiling rather neutral (Beveridge, Givens, Phillips, & Draper, 2009).

Findings from human face recognition studies also support the detrimental effect of incongruent emotional expressions on recognition accuracy. Bruce (1982), for example, manipulated both view (full face vs ¾ view) and emotional expression (smile vs neutral) in an old/new recognition paradigm and measured both recognition accuracy and latency (mean time to accept and reject 'old and 'new' faces respectively). The study reported 90% recognition hit rates for identical pictures which decreased to 81% when there was a mismatch in emotional expression (e.g. seeing a neutral image at learning and a smiling image at test). This pattern of results was later replicated using a 1 in 10 face matching task where participants saw a target image at the top and were asked to decide whether a different image of the same person was presented in an array of 10 other images (Bruce et al., 1999). There is also some evidence that faces with a smiling expression during face learning lead to higher recognition accuracy compared to faces displaying other emotional expressions (D'Argembeau, Van der Linden, Comblain, & Etienne, 2003; Shimamura, Ross, & Bennett, 2006) and positive affect in the face has been further shown to improve even familiar face recognition (Kaufmann & Schweinberger, 2004).

Given the findings of automatic recognition systems, it is surprising that no behavioural study has explored the influence of emotionally-

congruent face pairs on face matching performance. Experiments 10 and 11, therefore, aimed to address this issue and investigate whether a smile provides any further identity-diagnostic information that can be used to enhance face matching performance.

Method

Participants

A total of 40 participants (2 male, mean age = 19.6, age range = 19-24) from the University of York took part in the study. All had normal or corrected-to-normal vision and received payment or course credits for their participation. Informed consent was provided prior to participation and experimental procedures were approved by the ethics committee of the Psychology Department at the University of York.

Materials

A total of 180 images were used as experimental stimuli. These comprised four different images of 30 unfamiliar identities and two different images of 30 foil identities (see Figure 4.5 for examples). For each identity, there were two neutral and two smiling images paired with a neutral and a smiling foil image. In order to preserve consistency among the smiling images, only those with an open-mouth smile were used for this experiment. Match images were different images of the same identity and foil images were of a different identity matching the verbal description of the target identity. Both the neutral and smiling foil images paired with each identity were of the same person.

All images were downloaded from a Google Image Search by entering the name of the identity 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 as well as matching the emotional expression requirements. They were all "ambient" images that captured a good amount of face variability. The identities used in this experiment were non-UK professional athletes.

They were selected because they will be unfamiliar to UK viewers and because they will match the faces we encounter in everyday life better compared to faces of foreign celebrities. Celebrities have a specific way they behave in front of cameras – they smile professionally and have their make up and hair done by professional make up artists, which limits the variability captured in their photographs. Athletes, on the other hand, have a more natural behaviour and wear minimal make up.

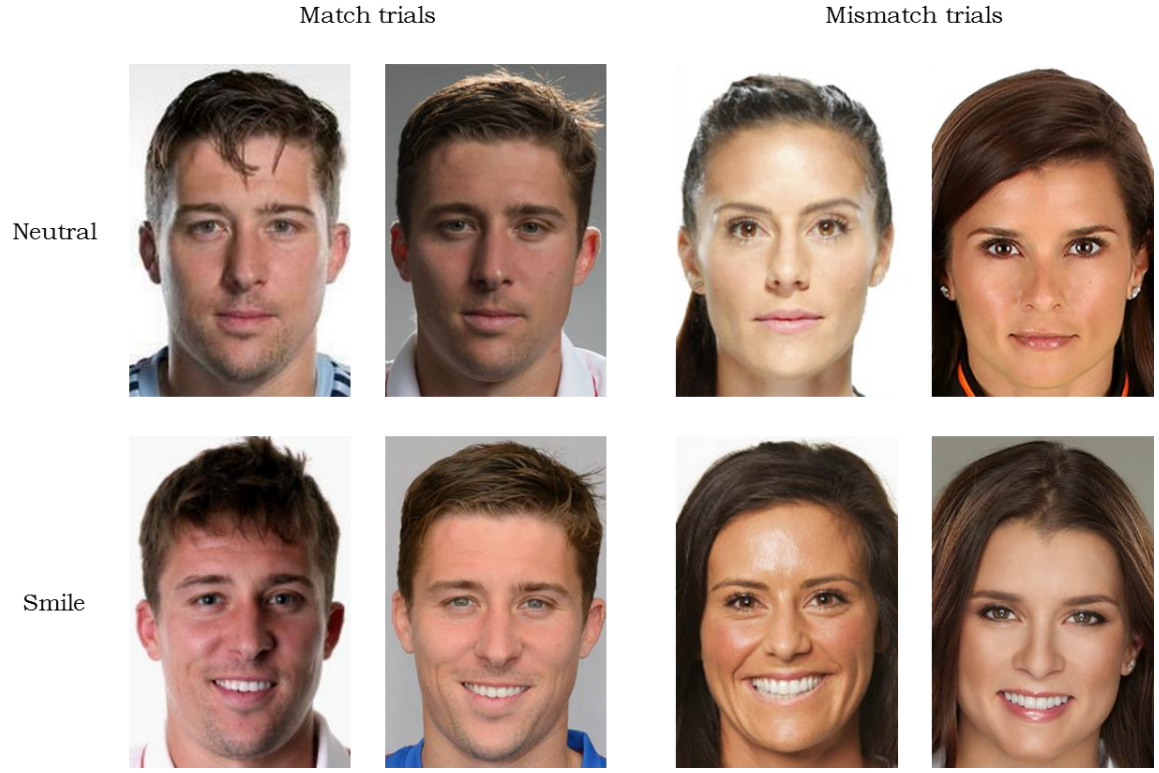

Figure 4.5. Trial type examples. Match trials used images of the same identity and mismatch trials used images of two different identities. For mismatch trials, each column contains images of the same identity.

Design and procedure

The experiment used a 2 (smile/neutral) x 2 (match/mismatch) design. All participants completed 60 trials of a face matching task. For this task, participants were presented with two images on the screen and asked to decide whether these images are of the same person or of two different people by pressing corresponding keys on the keyboard. The task was not timed,

however participants were encouraged to be as quick and accurate as possible. Participants completed an equal number of match and mismatch trials as well as an equal number of smile and neutral trials. They saw images of each identity twice, however the conditions they saw them in were counterbalanced so that participants never saw the same image twice. In match trials, participants were presented with two different images of the same identity, whereas in mismatch trials they saw images of two different identities (an image of the target identity and a foil image of an identity that matches the physical description on the target). Trials were also either neutral, where both images on the screen had a neutral expression, or smiling, where both images had a happy expression. Example match and mismatch trials across the two expressions can be seen in Figure 4.5. Trial order was randomised for each participant.

Results and discussion

Mean matching accuracy across all conditions is presented in Figure 4.6. A 2 x 2 within subjects ANOVA (expression: neutral vs smile; trial type: match vs mismatch) revealed a significant main effect of expression (F (1, 39) = 25.33, p < .001, η_p^2 = .39) as well as trial type (F (1, 39) = 24.31, p < .001, η_p^2 =.38). There was no significant interaction between expression and trial type (F (1, 39) < 1, p > .05, η_p^2 = .01).

Results showed that using smiling images in a matching task could improve performance in both match and mismatch trials. This is an important finding as most methods of improving face matching such as facial caricaturing (McIntyre, Hancock, Kittler, & Langton, 2013) or using multiple images per identity (White et al., 2014a) have had limited success. Here, we demonstrate a significant improvement for both match and mismatch trials by just providing further information about the face such as smile and teeth shape and smile lines around the mouth and eyes. This extends findings from automatic face recognition algorithms and shows a comparable effect for emotion-congruent face pairs in human performance. It is therefore possible that smiling provides further idiosyncratic information about people that makes it easier for them to be recognised.

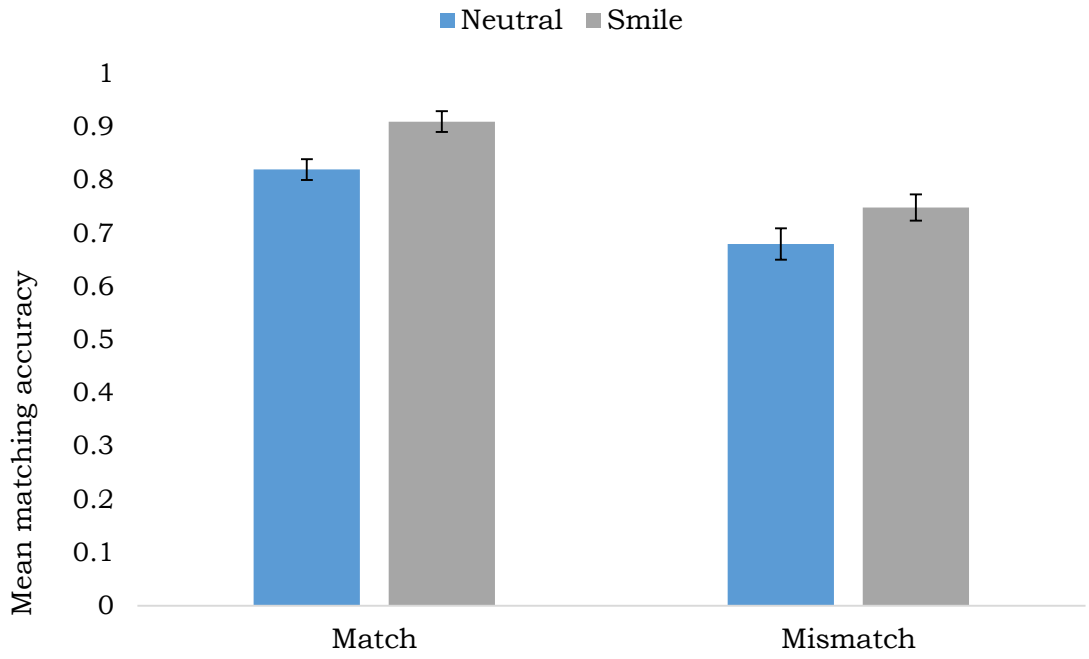

Figure 4.6. Mean matching accuracy across expression and trial type. Error bars represent within-subjects standard error (Cousineau, 2005).

4.5 Experiment 11

Introduction

The last experiment in this chapter aimed to follow up on the findings from Experiment 10 and explore the effect of smiling across different intensities. Here, we used the same neutral and smiling images from Experiment 10 but we also added an intermediate condition with images of the same identities displaying a closed-mouth smile. This way we will be able to detect whether the improvement can be achieved by any smile or the size and shape of teeth provide much of the identity-diagnostic information used to improve performance.

Method

Participants

A total of 60 participants (7 male, mean age = 20.6, age range = 19-27) from the University of York took part in the study. All had normal or corrected-to-normal vision and received payment or course credits for their participation. Informed consent was provided prior to participation and experimental procedures were approved by the ethics committee of the Psychology Department at the University of York. Only participants who had not taken part in Experiment 10 were recruited for the present experiment.

Materials

The same 30 identities as the ones in Experiment 10 were used for the present experiment. A further 3 images were collected for each identity – two images with a closed-mouth smile as well as an extra image of the same foil identity with a closed-mouth smile (see Figure 4.7 for examples). In order to ensure that the stimuli captured the desired emotional expression all images were rated by a separate sample of 54 participants. Participants were presented with all 270 images individually and asked to rate how happy the person in the image was on a scale from 1 (not at all) to 9 (extremely). Analysis was run by item rather than by participant. A one-way repeated measures ANOVA showed a significant main effect of expression (F (2, 58) = 607.41, p < .001, η_p^2 =.95). Follow-up Tukey HSD tests showed significant differences between all levels of the expression factor with open smiles (M = 6.97, SD = .34) rated as the happiest, followed by closed-mouth smiles (M = 5.45, SD = .46) and finally the neutral expression (M = 3.10, SD = .64). This validates the stimuli sample and shows clear differences in the intensity of emotional expressions across the three conditions.

Design and procedure

The experiment used a 3 (neutral / closed-mouth smile / smile) x 2 (match / mismatch) design. Other than the extra level of the expression factor, the experiment used the same design and procedure as Experiment 10. Participants completed 60 trials of the face matching task with an equal number of match and mismatch trials as well as an equal number of neutral, closed and open smile trials. Again, they saw images of each identity twice

but images were counterbalanced so that they never saw the same image twice. Examples of match and mismatch trials across all emotional expressions can be seen in Figure 4.7.

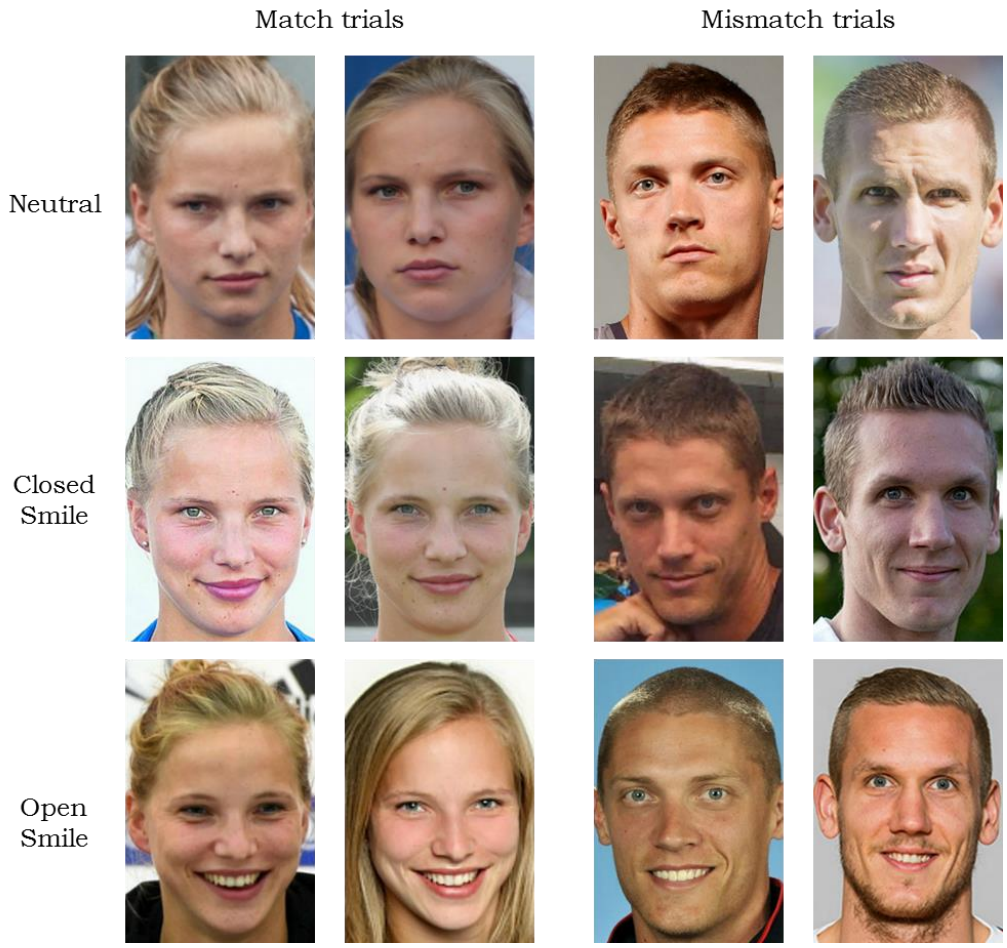

Figure 4.7. Trial type and stimuli examples for Experiment 11. Match trials used images of the same identity and mismatch trials used images of two different identities. For mismatch trials, each column contains images of the same identity.

Results and discussion

Mean matching accuracy across all conditions is presented in Figure 4.8. A 3 x 2 within subjects ANOVA (expression: neutral vs closed smile vs open smile; trial type: match vs mismatch) revealed a significant main effect of expression (F (2, 118) = 20.87, p < .001, η_p^2 = .26). There was no significant main effect of trial type (F (1, 59) < 1, p > .05, η_p^2 < .01) nor a significant

interaction between expression and trial type (F (2,118) = 2.80, p > .05, η_p^2 = .05). Follow-up Tukey HSD tests showed that face matching accuracy with smiling images was significantly higher than matching images with a neutral expression and a closed-mouth smile and that was true for both match and mismatch trials (p < .05). No difference in matching accuracy was found for images with a neutral expression and images with a closed-mouth smile.

These results replicate the findings reported in Experiment 10 that presenting participants with two smiling images improves their face matching accuracy for both match and mismatch trials. This further supports the idea that a smile might provide some additional information that is diagnostic of identity. No improvement was seen in the closed-mouth smile condition compared to the neutral condition and there were very clear differences in the intensity ratings of these two types of images. It seems that the perceptual information provided by the shape and size of the teeth as well as the distortion in the face produced by an open-mouth smile are more likely to drive the increase in accuracy for smiling images by providing further opportunity for the face to reveal more of its idiosyncratic features.

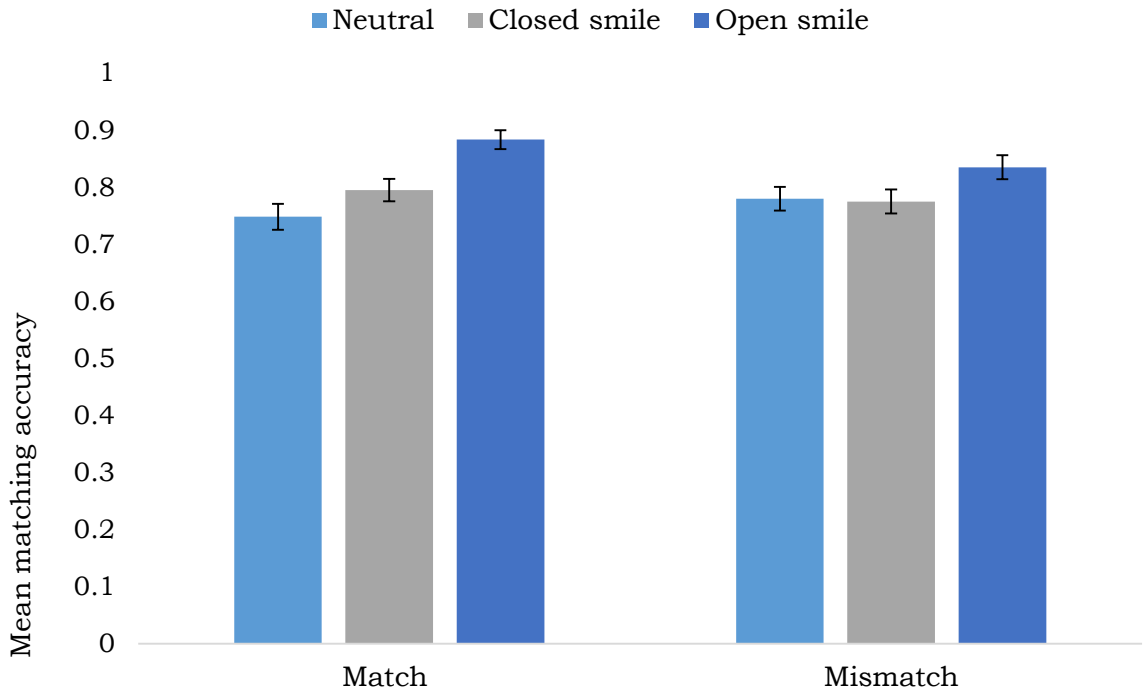

Figure 4.8. Mean matching accuracy across expression and trial type conditions. Error bars represent within-subjects standard error (Cousineau, 2005).

4.6 General Discussion

The experiments in this chapter aimed to investigate the association between social attribution and face matching performance. These were rather exploratory than theoretically grounded experiments based on evidence for the facilitatory role of social judgements on face memory (Courtois & Mueller, 1979; Winograd, 1976). Furthermore, both identity recognition and social evaluation are affected by the same factors including world and person variability (e.g. lighting, emotional expressions). When comparing ratings of images used in match and mismatch trials, there were clear differences in attractiveness ratings, such that images of the same person received more similar attractiveness ratings than images of two different people. Relating attribute differences to matching performance, however, indicated weak, if any, correlations between social attribute differences and matching accuracy for both match and mismatch trials. Nevertheless, there was a significant relationship between attractiveness and matching performance on mismatch trials, indicating that participants were more likely to correctly identify images as belonging to two different identities with a larger difference in their attractiveness ratings. The following two experiments focused on a specific social evaluation factor – emotional expression - and investigated whether smiling can provide additional identity information. Results showed a significant increase in face matching accuracy when images in each face pair displayed a smile rather than a neutral expression, demonstrating the potential of smiling images to maximise matching performance.

The overall findings of no significant correlations between social attribute differences and matching performance (apart from attractiveness in mismatch trials only) is inconsistent with findings from studies using face memory tasks (Bower & Karlin, 1974; Courtois & Mueller, 1979). This implies that the deeper level of processing brought about by social evaluation is associated with an improvement in face memory, but not perceptual tasks such as matching. Moreover, the different pattern of results for match and mismatch trials suggests that participants might be activating different cognitive processes and engaging in different strategies for their identity decisions. It is therefore possible that once they have formed an initial

hypothesis about the outcome of the trial, they are influenced by the evidence that confirms it. As images of the same person are more likely to lead to a match response and images of two different people are more likely to lead to a mismatch response, it might be that participants are focusing on similarities in mismatch trials and on differences on mismatch trials. Thus, it is not surprising that we find a significant association between attribute differences and matching accuracy for mismatch trials only. It is also interesting to note that comparing attractiveness ratings of images used in match and mismatch trials shows that people can actually detect differences in identity as images of the same person received more similar ratings than images of two different people. This information, however, seems to be used in mismatch trials only where cues to differences might be of a greater relevance.

Embedding images in a passport frame replicated the findings from the first experiment. It also demonstrated the effect of context as there was an additional significant correlation between dominance and matching accuracy. This is somewhat surprising as a passport control situation might be expected to activate the importance of trustworthiness, rather than dominance. As models of threat perception, however, describe trustworthiness as someone's intent to do harm and dominance as their ability to do harm (Oosterhof & Todorov, 2008), it is possible that dominance is the more important and relevant cue in this context. Moreover, results showed a different pattern of results for match and mismatch trials. This implies that different processes are involved when looking for reasons to classify two images as being of the same identity and when looking for reasons to classify two images as being of two different identities.

Findings from the present chapter can also be related to the results from Chapter 3 where we see a lot more between- rather than within-person variability for ratings of attractiveness. This is consistent with our findings here that participants used this trait in particular when examining images of two different identities in mismatch trials and supports the idea that different images of the same person can give rise to different social evaluation ratings when it comes to trustworthiness and dominance, but not attractiveness. As

attractiveness ratings were based on identity to a greater extent it is possible that differences in this social evaluation dimension are used as a more reliable cue to differences in identity. Trustworthiness and dominance, on the other hand, are affected by the physical properties of images to a greater extent which could make them unreliable for identity-based decisions.

The last experiments in the chapter focused on the influence of emotional expressions and smiling, in particular, on face matching performance. This was motivated by the possibility that smiling faces might present participants with some extra identity-diagnostic information from the shape of the smile and teeth as well as wrinkles around the mouth and eyes. Results provided support for this suggestion, showing higher matching accuracy when both images in the face pair had a smiling rather than a neutral expression. Such findings are consistent with automatic face recognition studies which show that smiling images are much better recognised than neutral images (Beveridge et al., 2009; Yacoob & Davis, 2002). While face recognition algorithms might not necessarily simulate the exact processes of human face recognition, our results demonstrated that people are actually able to extract the information provided through a smile and use it in a constructive way to improve recognition rates. It should be noted that findings from these experiments are not in contrast to human recognition studies demonstrating a significant decrease in performance with the introduction of expression incongruence (Bruce, 1982; Bruce et al., 1999). These studies explore a different key comparison – while they compare trials where one image has a neutral expression and the other a smiling expression, the present studies investigated congruent pairs only (i.e. both images in the face pair have either a smiling or a neutral expression).

What is probably most impressive about the improvement in matching performance, brought about by a smile, is that this advantage was seen in both match and mismatch trials. This implies that smiling can overcome differences in match and mismatch mechanisms and provide identity-diagnostic information that is relevant both in situations where we need to compare images of the same person and images of different people. This is in

contrast to most methods of improving matching performance established so far, such as feedback which has been shown to improve performance on mismatch trials only or within-person variability that improves performance on match trials only (Alenezi & Bindemann, 2013; White et al., 2014a).

The experiments in this chapter demonstrated that participants can detect differences in identity as seen through their attractiveness ratings of images belonging to the same person and images belonging to two different people. This information, however, was only utilised in mismatch trials where participants might be more likely to look for differences rather than similarities. Nevertheless, the majority of non-significant correlations between social attributes and matching performance leads to the conclusion that social cues are not highly relevant in perceptual identity tasks. Relating these findings back to previous literature on the relationship between social attribution and face memory, demonstrates that the beneficial effect of evaluating faces on social dimensions applies to face memory but not to perceptual identity decisions. We also identified a successful way of improving face matching accuracy making use of emotional expressions and the information extracted from a smile, in particular. Results showed that image pairs with a smiling rather than a neutral expression led to a more accurate face matching performance, possibly due to the additional information provided by a smile (e.g. shape of smile and teeth, smile lines around the mouth and eyes). This was true for both match and mismatch trials making this approach superior to other matching improvement strategies such as feedback or using multiple images per identity which have been shown to increase accuracy in only one those matching components. Overall, our findings showed that conceptual information such as attributing social ratings to a face might be more likely to improve performance on face memory tasks as it provides further depth of processing. Perceptual information such as the one provided by emotional expressions, on the other hand, is more useful for memory-independent identity tasks such as face matching.

Chapter 5 – Audiovisual Integration in First Impressions

5.1 Introduction

A wealth of biological and social information about people, such as sex, age, ethnicity or emotional state, can be inferred by either looking at their faces or listening to their voices (Belin et al., 2011; Bruce & Young, 1986; Yovel & Belin, 2013). Moreover, we constantly recognise people's identities from their faces and voices, for example by looking at a photograph or hearing a voice on the telephone. People infer socially-relevant information and form stable first impressions about unfamiliar others from both faces and voices (Todorov et al., 2009; Zuckerman & Driver, 1989). Social impressions from faces arise very quickly (after less than a second of exposure in many reports), whereas impressions from voices will always include some temporal element.

Parallels between face and voice social perception encompass their structure, consistency and implications. Using the same approach as Oosterhof and Todorov (2008), McAleer et al. (2014) demonstrated a two-dimensional space for social evaluation of voices with valence and dominance as the main dimensions. Such findings are consistent with the trustworthiness/dominance face model as well as other social evaluation models such as concept evaluation (Osgood et al., 1957), group evaluation (Fiske et al., 2007) and models of interpersonal perception (Wiggins, 1979), all of which rely on two orthogonal dimensions - affiliation and dominance. While first impressions might not represent reality accurately, social evaluation is characterised by a high level of agreement between observers or listeners for both facial and vocal information (McAleer et al., 2014; Zebrowitz & Montepare, 2008). This implies that people use consistent physical information in the face and acoustic information in the voice to inform their social judgements. Furthermore, zero-acquaintance impressions from voices present with comparable social implications, with studies demonstrating that voting outcomes can be predicted not only by the perceived competence in the face (Ballew & Todorov, 2007; Olivola & Todorov, 2010a) but also by the pitch

of the voice (Tigue et al., 2012). Similarly, both facial and vocal information have been shown to predict courtroom outcomes (Chen et al., 2016; Wilson & Rule, 2016) as well as to influence dating and mate preferences (Little et al., 2006; Wells, Dunn, Sergeant, & Davies, 2009).

Audiovisual integration

In this chapter, we aim to explore first impressions gained from multimodal stimuli, comprising faces and voices. Given that both these sources individually have been shown to give rise to consistent social attributions, how do they interact? Do voices or faces dominate in social judgements, or does the signal from one source influence the interpretation of the other? Strong integrative effects have already been shown in speech perception (McGurk & MacDonald, 1976; Summerfield, 1979), identity recognition (Ellis et al., 1997; Schweinberger et al., 1997), and emotion classification (Hess et al., 1988; Mehrabian & Ferris, 1967). Person identification studies, for example, show that participants are generally quicker and more accurate when identifying people from their faces, rather than their voices. Priming studies, however, highlight the importance of the voice demonstrating that participants are quicker to identify a face as familiar after being presented with the voice of that same identity and vice versa (Schweinberger et al., 1997). Visual information from the face has also been shown to be more critical in emotion classification. De Gelder and Vroomen (2000), for example, paired morphs of the Ekman faces (Ekman & Friesen, 1976), going from a happy to a sad expression, with voice recordings of sentences pronounced in a happy or sad way. Their results showed that while both face and voice cues contributed significantly to emotion classification, the face had a much stronger effect.

Further, there is evidence that audio-visual integration in emotion recognition is an automatic process as participants seem to incorporate face and voice cues together, even when they are instructed to ignore one of the information channels. De Gelder and Vroomen (2000) found a significant effect for both the visual and vocal channels on the perception of happiness/sadness and happiness/fear when participants were presented

with both channels but specifically instructed to ignore either the face or the voice when making their judgements. Evidence for the automatic nature of audio-visual integration also comes from studies on identity recognition (Campanella & Belin, 2007). In a series of experiments Schweinberger et al. (2007, 2011) demonstrated that presenting participants with corresponding and non-corresponding face-voice pairs had an influence on familiarity decisions: recognition of a familiar voice was faster and more accurate when it was paired with the corresponding face – even when participants were specifically instructed to make their judgements exclusively based on the audio cues.

In comparison with research examining emotion and identity recognition from faces and voices, comparatively fewer studies have explored the effect of combining visual and vocal cues on the formation of first impressions. This is in spite of features such as dominance, trustworthiness, and attractiveness forming a key part of prominent social perception models (Fiske et al, 2007; Oosterhof & Todorov, 2008). Rezlescu et al (2015) examined listener perceptions of attractiveness, trustworthiness, and dominance using a combination of static male faces and brief vowel sounds, produced by male speakers adopting a variety of emotional vocal expressions such as happy, sad, and angry. The results indicated that facial information was more influential in judgements of attractiveness, whereas vocal information was more influential in dominance judgements. Both visual and vocal information contributed significantly to trustworthiness judgements. However, Tsankova et al. (2015) examined perceptions of trustworthiness using facial and vocal cues and argued that trustworthiness judgements were more heavily influenced by facial rather than vocal information.

Natural face and voice variability

In order to address our limited understanding of the combined effects of vocal and facial cues on social evaluation, this chapter aims to investigate the relative contribution of audio and visual information to the perception of the fundamental social perception dimensions – trustworthiness and dominance. We also aim to explore whether this audio-visual integration is

automatic, extending our knowledge about integrated person perception. Our approach differs from that taken in previous studies in that we use vocal stimuli comprising speech, which (arguably) represent real-world social interactions more accurately than non-verbal vocalisations. While some argue that the use of brief, neutral vowel sounds mitigates the influence of aspects of voice such as prosody and semantic content (Rezlescu et al., 2015), the extent to which this replicates real everyday speech has been the topic of debate (Apple et al., 1979). Social evaluations are clearly multi-faceted in everyday life, and so there is some value in studying them using contentful utterances.

We also make use of within-person variability to manipulate these social evaluations. In most studies of first impressions, it is assumed that *people* give rise to stable judgements, i.e. a particular person is more or less trustworthy, dominant etc. However, this is now known to be false. Ratings for different photos of the same person can vary more than for photos of different people (Jenkins et al., 2011; Todorov & Porter, 2014, also demonstrated in Experiments 3 & 4). First impressions derived from faces can therefore reflect differences in *photos* rather than differences in *people*. Instead of using different identities rated as high or low in dominance and trustworthiness, here we sample different images of the same identity and select those rated as the most and least trustworthy and dominant. We also isolate the effect of a single acoustic measure – mean pitch – which has previously been linked to perceptions of dominance and trustworthiness in voices (Ohala, 1984; Tsanani et al., 2016).

Overview of studies

In Experiment 12, we first validate a set of vocal stimuli and investigate the role of pitch in dominance perception. In Experiment 13, these auditory stimuli were matched with a set of face images perceived as high and low in dominance to investigate the relative effects of both channels on social person perception. Experiment 14 extends work on the automaticity of audio-visual integration (de Gelder & Vroomen, 2000; Schweinberger, 2007) into the domain of first impressions. We present participants with both facial and

vocal cues and instruct them to ignore one of those channels when they evaluate each person. Experiments 12-14 focus on the perception of dominance and Experiments 15 and 16 extend these into the perception of trustworthiness. Experiment 15 evaluates the use of pitch as a cue for trustworthiness, and Experiment 16 examines multimodal trustworthiness perception.

5.2 Experiment 12

Introduction

This first experiment was conducted to obtain baseline judgements of dominance for our auditory stimuli, independent of visual information. The specific vocal parameter investigated in this experiment is mean fundamental frequency (F0), which we label mean pitch following Laver's (1994) assertion that the two terms can be used interchangeably in spite of a strictly non-linear relationship. We manipulated the pitch of vocal stimuli, hypothesising that this would affect perception of dominance. Pitch has been highlighted as one of the most perceptually salient acoustic cues used by listeners to infer emotion and affect in speech (Dimos et al., 2015). Following work which identifies low pitch as a signal of aggression and dominance across a variety of animal species (Morton, 1977), research has identified a perceptual link between the lowering of F0 and the perception of both social and physical dominance in human speech (Ohala, 1984; Puts et al, 2006, 2007; Tusing & Dillard, 2000).

It is important to establish whether pitch manipulation has the hypothesised effect in verbal stimuli from male and female speakers. Previous literature is somewhat contradictory, perhaps reflecting the wide diversity in the types of stimuli used (Borkowska & Pawlowski, 2011; McAleer et al., 2014; Tsantani et al., 2016; Vukovic et al., 2011). To anticipate the results, we found that verbal utterances were judged more dominant when rendered in lower pitch – an effect which held for both male and female voices.

Method

Participants

Voices were rated by 36 participants (13 male, mean age = 23.9, age range = 18-36). All participants were students at the University of York and received payment or course credits for their participation. Informed consent was provided prior to participation in accordance with the ethical standards stated in the 1964 Declaration of Helsinki.

Materials

Experimental stimuli were 40 voice recordings (2 for each of 20 identities, one manipulated to a higher pitch and the other manipulated to a lower pitch). Twenty speakers (10 male, mean age = 23, age range = 18-35) gave informed consent to be recorded producing the utterance *"I wouldn't do that if I were you"*. Voices were recorded following ethical consent from the Department of Language and Linguistic Science at the University of York. All speakers were students at the University of York. Recordings were conducted in quiet recording environments using a Zoom H4N handheld recorder, with the built-in microphone positioned 30cm from each speaker.

The utterance *"I wouldn't do that if I were you"* was chosen due to its indirect nature (Searle, 1979) and because it can give rise to a range of social inferences – including interpretations that it represents advice or threat. Our approach therefore differs from those based on presentations of neutrally-worded reading passages or on non-verbal vocalisations (e.g. vowels sounds), which are very commonly used in this field (Berry, 1991; Rezlescu et al., 2015). Digital manipulations using Praat (Boersma and Weenink, 2016) were used in order to create contrasting mean pitch levels for each stimulus. A Praat pitch alteration script (Fecher, 2015) was used to create low and high mean pitch levels. For male speakers, the mean F0 of each recording was altered to 90Hz (low) and 140Hz (high). These values are 25Hz above and below an approximation of an average male mean F0 level (Hudson, De Jong, McDougall, Harrison, & Nolan, 2007; Künzel, 1989; Lindh, 2006), and

represent values in the highest and lowest 10% of population values reported by Hudson et al. (2007). For female speakers, the mean F0 of each recording was altered to 170Hz (low) and 250Hz (high). These values are 40Hz above and below an approximation of an average female F0 level, and reflect the low and high ends of the mean F0 range reported for female speakers (Künzel, 1989; Traunmüller & Erickson, 1995). All recordings were checked to ensure that no digital artefacts had influenced the sound quality as a result of the editing process.

Procedure

Data were collected online using Qualtrics software (2015, Provo, UT) given that prior research on online and lab-based samples finds the two comparable in terms of means, standard deviations and internal reliability (Germine et al., 2012; Horton, Rand, & Zeckhauser, 2011). Participants were presented with each recording individually and asked to rate dominance on a scale from 1 (not at all dominant) to 9 (extremely dominant). Participants rated all 40 of the vocal stimuli, each in an independently randomised order.

Results and discussion

Dominance ratings had very high inter-rater reliability (Cronbach's α = .89). A paired t-test showed that low-pitched voices (M = 4.82, SD = 1.05) were perceived as significantly more dominant than high-pitched voices (M = 3.80, SD = 1.09), t (35) = 6.81, p < .001, d = 1.13. This is consistent with previous studies investigating the effect of vocal pitch on the perception of dominance and aggression (Ohala, 1984).

Despite an overall effect of pitch on perceived dominance, some work with different types of stimuli has suggested that such effects are modulated by speaker gender (McAleer et al., 2014; Tsantani et al., 2016). This was not the case for our stimuli, which showed a consistent effect of pitch manipulation for both male speakers (Means: 4.39 vs 5.31; t (35) = 4.87, p < .001, d = .81) and female speakers (Means: 3.22 vs 4.34; t (35) = 5.94, p < .001, d = .99).

Having established that the pitch manipulation has the hypothesised effect – i.e. that it is possible to make the same voice sound more or less dominant – we now progress to multimodal experiments in which we combine faces and voices.

5.3 Experiment 13

Introduction

In this experiment, we use the vocal recordings validated in Experiment 12, and pair them with a set of facial stimuli, in order to explore how face and voice evaluations come together to form an integrated impression of dominance. Rezlescu et al. (2015) report that when participants were required to make dominance judgements to multimodal stimuli (face-voice), their judgements were more influenced by the voices than the faces (a pattern which was reversed for ratings of attractiveness). Experiment 13, therefore, builds on this finding, but with the following differences.

First, our manipulations of stimulus dominance are not confounded by identity. So, here we present high and low-dominance versions of *the same voices*, as prepared by the pitch manipulation described in Experiment 12. We also present high and low-dominance versions of *the same faces* by picking images which have been rated independently. Second, this experiment uses voices articulating verbal speech, as described in Experiment 12. Participants hear the same phrase uttered across all combinations of conditions, rather than hearing the content-free vocalisations of some earlier studies. This has the advantage that the speech signal is meaningful – while avoiding any confounding of condition with content.

To anticipate the results, we found additive effects of face and voice on overall judgements of dominance. Dominance of both faces and voices independently contributed to the impression formed when stimuli were presented multimodally. However, consistent with Rezlescu et al (2015) we found that voices had the larger effect on overall judgements.

Method

Participants

60 participants (16 male, mean age = 21.9, age range = 18-32) took part in the experiment. All were students at the University of York. All participants had normal or corrected-to-normal vision, reported no hearing impairments and received payment or course credit for their participation. Informed consent was provided prior to participation and experimental procedures were approved by the ethics committee of the Psychology Department at the University of York.

Design

The experiment used a 2 (face/voice) x 2 (high/low dominance) design. All participants completed 40 trials (10 per condition) in which a face and a voice was presented together, meaning that over the session, participants saw two different images of each stimulus person's face, and heard two different versions of each stimulus person's voice. Across the experiment, trials were counterbalanced such that all combinations of high-/low-rated faces and voices were presented equally often. Trial presentation order was randomised independently for each participant.

Materials

Voice recordings from Experiment 12 were used as audio stimuli. Face stimuli were selected from the 20-20 set from Experiment 3. It included 400 images comprising 20 images each of 20 unfamiliar identities downloaded from an internet search. All images were highly variable or 'ambient' (Jenkins et al., 2011) and therefore captured a great amount of variability within each identity due to different lighting conditions, emotional expressions, pose, etc. (see Figure 3.1 for examples).

For the purposes of the present experiment, we selected the images that were rated as the most and least dominant for each identity. This

provides sets of 20 high- and 20 low-dominance images, with the same identities in each set. Paired t-tests confirmed that images in the high dominance group (M = 6.47, SD = .62) were perceived as significantly more dominant than those in the low dominance group (M = 4.05, SD = .55, t (19) = 17.48, p < .001, d = 3.94). Figure 5.1 shows examples of those images for a male and female identity.

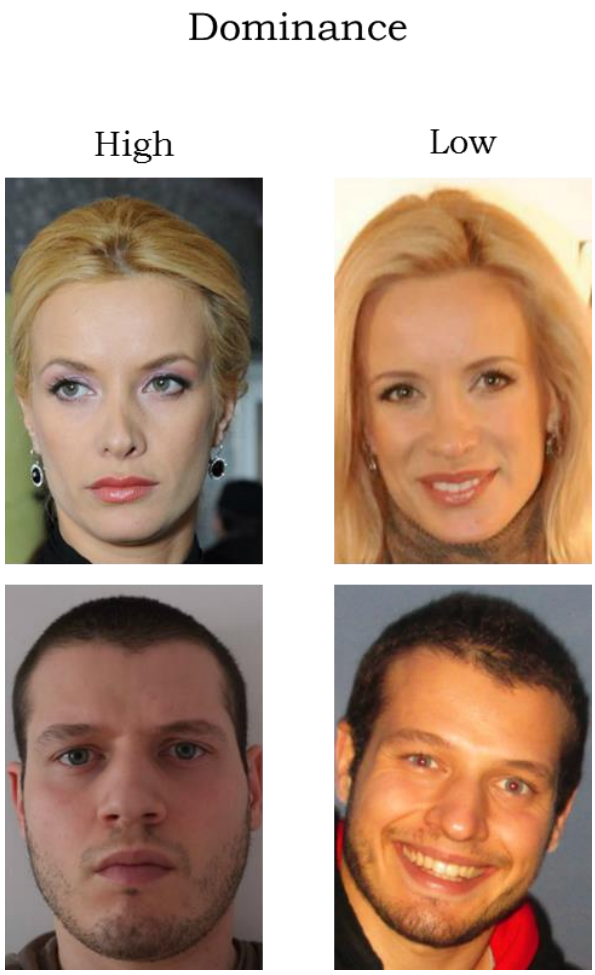

Figure 5.1. Different images of the same people rated as high and low in dominance.

Procedure

Each trial comprised a face and a voice presented simultaneously. The vocal stimuli played automatically through closed-cup headphones and were presented once only. Participants' task was to rate each identity for dominance on a scale from 1 (not at all dominant) to 9 (extremely dominant).

Face stimuli were presented on a white background at the centre of the screen and the rating scale was positioned below the face image. Participants indicated their response by pressing the corresponding key on the keyboard. The task was not timed, and participants were given no further definition of 'dominance', but encouraged to rely on their 'gut feeling' (Oosterhof & Todorov, 2008).

Results and discussion

Mean ratings by condition are shown in Table 5.1. A 2 x 2 within subjects ANOVA revealed significant main effects of face dominance (F (1, 63) = 72.23, p < .001, η_p^2 = .53) and voice dominance (F (1, 63) = 250.92, p < .001, η_p^2 = .80), with no interaction (F (1, 63) < 1, p > .05, η_p^2 = .01).

Table 5.1. *Mean Ratings of Dominance across Conditions in Experiment 13. SDs in Parentheses.*

	Low dominance voice	High dominance voice
Low dominance face	4.0 (.46)	5.1 (.58)
High dominance face	4.6 (.47)	5.8 (.47)

Our results show clear, independent contributions of face and voice on dominance judgements for multimodal stimuli. Interestingly, the two sources of information do not interact, but provide completely additive contributions to the overall judgement. This is consistent with the findings of Rezlescu et al. (2015) who found no correlations between judgements of dominance on the faces and voices of the same people – providing compelling evidence against the validity of these attributions, despite their strong consensus (as replicated here). We also show a similar effect of information source as Rezlescu et al. (2015). While both face and voice predict overall dominance ratings, the voice manipulation produces a larger effect. This is consistent with earlier findings on the importance of auditory information for the perception of dominance

and aggression and could be explained with its higher reliability. Dominance judgements have been shown to correlate highly with sexually dimorphic aspects, and vocal pitch is a sexually dimorphic aspect of voice (Puts et al., 2006). Vocal pitch might, therefore, be a more reliable channel when assessing someone's masculinity, which is related to dominance (Collignon et al., 2008).

Our results suggest a rather straightforward, additive system of audiovisual integration for the perception of dominance. Two questions therefore arise. In the following experiment we ask how automatic is this process, i.e. to what extent can one weigh either source of evidence through top-down control. Following this, we then return to first impressions more generally, and ask whether this same pattern of additive effects exists for the other fundamental dimension of social evaluation – trustworthiness.

5.4 Experiment 14

Introduction

In the study of emotion perception, there is clear evidence that cues from voices and faces are combined to some extent in a mandatory way. For example, when presented with multimodal stimuli (face and voice) and asked to make a judgement about the person's emotional state, participants incorporate both voice and face cues, even when instructed to base their judgements on just one of these sources (de Gelder & Vroomen, 2000). In this experiment, we ask whether there is similarly a level of automaticity in cue combination when making judgements of dominance – i.e. making a social judgement rather than an emotional one. To do this, we replicate Experiment 13, but this time instruct participants to base their judgements on just one of the cues, voices or faces. If they are able to ignore a competing cue from another channel, this will provide evidence against mandatory combination of cues. To anticipate the results, we find evidence in favour of some mandatory cue combination – based on the result that participants' judgements are consistently influenced by the cues they are instructed to ignore.

Method

Participants

80 participants (8 male, mean age = 19.6, age range = 18-32) from the University of York took part in the experiment. All had normal or corrected-to-normal vision, reported no hearing impairments and received payment or course credit for their participation. Informed consent was provided prior to participation and experimental procedures were approved by the ethics committee of the Psychology Department at the University of York.

Design and procedure

The experiment followed exactly the same procedure as Experiment 13, using the same materials. As above, participants were shown 40 multimodal stimulus trials (face and voice), and asked to make a judgement of the person's dominance. However, in this case half the participants were instructed to make their judgements based on the face only, and the other half to make their judgements on the voice only. Participants were allocated to the two groups at random, and all other counter-balancing and trial sequence randomisation was the same as in Experiment 13.

Results and discussion

Mean ratings by condition are shown in Figure 5.2. A three-way mixed-design ANOVA (Instructions: focus on face vs voice; high vs low face dominance; high vs low voice dominance) showed significant main effects of face type (F (1, 78) = 185.29, p < .001, η_p^2 = .70) and voice type (F (1, 78) = 193.71, p < .001, η_p^2 = .71), but no significant three-way interaction, (F (1, 78) = 1.22, p > .05, η_p^2 = .02). Although we did not find a significant main effect of instructions (F (1, 78) < 1, p > .05, η_p^2 = .01), two-way interactions between instructions and face type (F (1, 78) = 69.52, p < .001, η_p^2 = .47) and instructions and voice type (F (1, 78) = 83.14, p < .001, η_p^2 = .52) were both significant. Across the instruction conditions face type had a much stronger effect when participants were instructed to focus on the face (F (1, 78) = 240.90, p < .001, η_p^2 = .76) than when they were instructed to focus on the

voice (F (1, 78) = 13.91, p < .001, η_p^2 = .15). The same pattern was observed for the effect of voice type – it was much stronger when participants were instructed to focus on the voice (F (1, 78) = 265.33, p < .001, η_p^2 = .77) than when they were instructed to focus on the face (F (1, 78) = 11.52, p < .01, η_p^2 = .13) showing that participants followed the instructions of the experiment. More importantly, the channel that participants were instructed to ignore nevertheless had a significant effect on their dominance ratings demonstrating that audio-visual integration is an automatic process that can be controlled to some, but not complete extent.

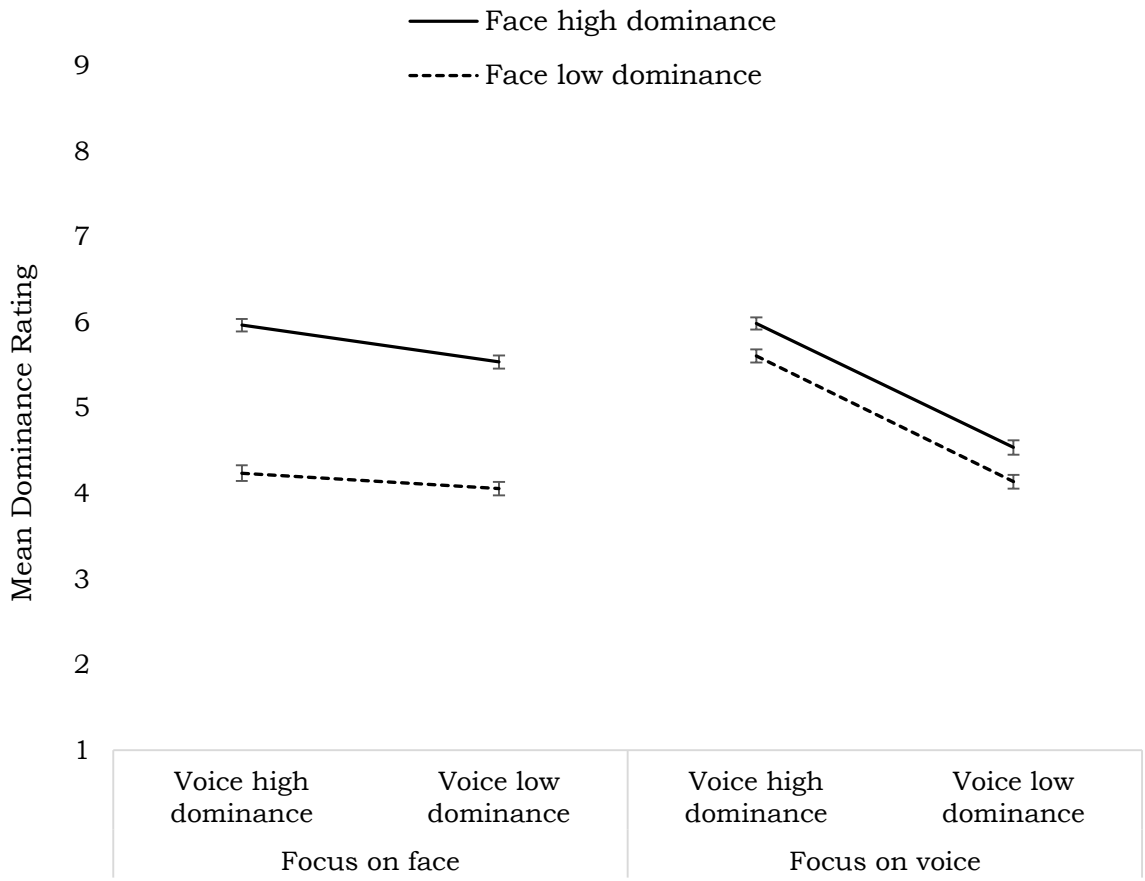

Figure 5.2. Mean dominance ratings for face-voice pairings under different instructions. Error bars are within-subjects standard error (Cousineau, 2005).

These results show two interesting effects. First, the instructions clearly influenced participants' behaviour. When instructed to focus on faces, the face type had the largest effect on dominance ratings. Similarly, when instructed to focus on voices, the voice type had the largest effect on ratings. Second, and despite this, the cue which participants were instructed to ignore, nevertheless had a significant effect on dominance ratings in each case. Furthermore, the effect was independent of the attended cue – there was no significant interaction between attended and ignored cue in either case. These results provide quite clear evidence for some degree of automaticity in the combination of multimodal information in social judgements of dominance. It would appear that the pattern reported in previous work for multimodal perception of identity and emotions (de Gelder & Vroomen, 2000; Schweinberger et al., 2007), also holds for social evaluation.

So far, we have concentrated primarily on the perception of dominance. We have shown that this attribution is made by independent contributions from voices and faces, and there is some degree of mandatory combination of these. In the next two experiments, we examine a different social judgement, trustworthiness. We ask whether the pattern of multimodal combination is the same for this judgement as it is for perception of dominance.

5.5 Experiment 15

Introduction

In Experiment 12, we demonstrated that pitch manipulation affects the perception of dominance in voices making verbal utterances. In order to study the multimodal perception of trustworthiness (Experiment 16), we first need to establish whether a simple voice manipulation gives rise to reliable changes in perception of this dimension. In fact, there are some reasons to believe that simple pitch manipulation will alter perception of trustworthiness, as it does for dominance. For example, Tsantsani et al. (2016) report a tendency for hearers to judge lower-pitched voices as more

trustworthy, both in male and female voices, albeit for temporally reversed speech. However, Vukovic et al. (2011) found no effect of pitch on trustworthiness judgements. Here we examine whether the voice samples used in Experiment 12 – in which pitch is raised or lowered for a spoken sentence - will also give rise to differences in trustworthiness judgements.

Method

Participants

Voices were rated by 38 participants (10 male, mean age = 21.6, age range = 18-35). All participants were students at the University of York and received payment or course credits for their participation. Experimental procedures were approved by the ethics committee of the Department of Language and Linguistic Science at the University of York.

Materials and procedure

Experimental stimuli were the same 40 voice recordings as used for Experiment 12, i.e. 2 for each of 20 identities, one manipulated with a higher pitch and the other manipulated with a lower pitch. Once again, data were collected online using Qualtrics software. Participants were presented with each recording individually and asked to rate it for trustworthiness on a scale from 1 (not at all trustworthy) to 9 (extremely trustworthy). The order of stimuli was randomised independently for each participant.

Results and discussion

Trustworthiness ratings had very high inter-rater reliability (Cronbach's α = .93). However, we there was no difference between trustworthiness ratings for high- (M = 5.08, SD = .61) and low- (M = 5.00, SD = .60) pitched voices (t (19) = 1.07, p > .05, d = .25), regardless of speaker gender. On this basis, we cannot use manipulated versions of the same voice in order to study multimodal perception of trustworthiness. For this reason, in the final experiment, below, we selected natural stimulus voices which had been independently rated as being high or low in trustworthiness.

5.6 Experiment 16

Introduction

In this final experiment, we replicated the approach taken in Experiment 13 by presenting participants with face-voice parings, and asking them to judge the trustworthiness of the person depicted. Faces and voices, which had previously been rated as high or low in trustworthiness, were presented in all combinations (high/low face/voice). To anticipate the results, we found independent effects of face and voice trustworthiness, with ratings being influenced more by facial rather than vocal cues.

Method

Participants

40 participants (8 male, mean age = 20.1, age range = 18-30) took part in the experiment. All were students at the University of York. All participants had normal or corrected-to-normal vision, reported no hearing impairments and received payment or course credit for their participation. Informed consent was provided prior to participation and experimental procedures were approved by the ethics committee of the Psychology Department at the University of York.

Design

The experiment used a 2 (face/voice) x 2 (high/low trustworthiness) design. All participants completed 40 trials (10 per condition) in which a face and a voice were presented together, meaning that over the session, participants saw two different images of each stimulus person's face, and heard two different versions of each stimulus person's voice. Across the experiment, trials were counterbalanced such that all combinations of high-/low-rated faces and voices were presented equally often. Trial presentation order was randomised independently for each participant.

Materials

The voice recordings from Experiment 15 were used as audio stimuli. We performed a median split on ratings of trustworthiness, separately for male and female voices. Combining male and female voices into high and low trustworthy groups gave means of 5.48 and 4.61 respectively (*SD*s = .33 and .47), a highly reliable separation (t (19) = 12.05, p < .001, d = 2.96). Note, that the results of Experiment 15 require that identities are no longer unconfounded with voice stimulus dimension. The high- and low-rated stimulus groups contain some voices of the same people – albeit manipulated to different pitches.

Face stimuli come from the same database as used in Experiment 3 (20 images of 20 people). To create high and low trustworthy groups, we selected the image for each individual which received the highest and lowest mean ratings. Figure 5.3 shows examples for a male and female identity. Paired t-tests confirmed that images in the high trustworthiness group (M = 6.29, SD = .46) were perceived as significantly more trustworthy than those in the low trustworthiness group (M = 4.58, SD = .46), t (19) = 15.69, p < .001, d = 3.51.

Procedure

Each trial comprised a face and a voice presented simultaneously. The vocal stimuli played automatically and were presented once only. Participants' task was to rate each identity for trustworthiness on a scale from 1 (not at all trustworthy) to 9 (extremely trustworthy). Face stimuli were presented on a white background at the centre of the screen and the rating scale was positioned below the face image. Participants indicated their response by pressing the corresponding key on the keyboard.

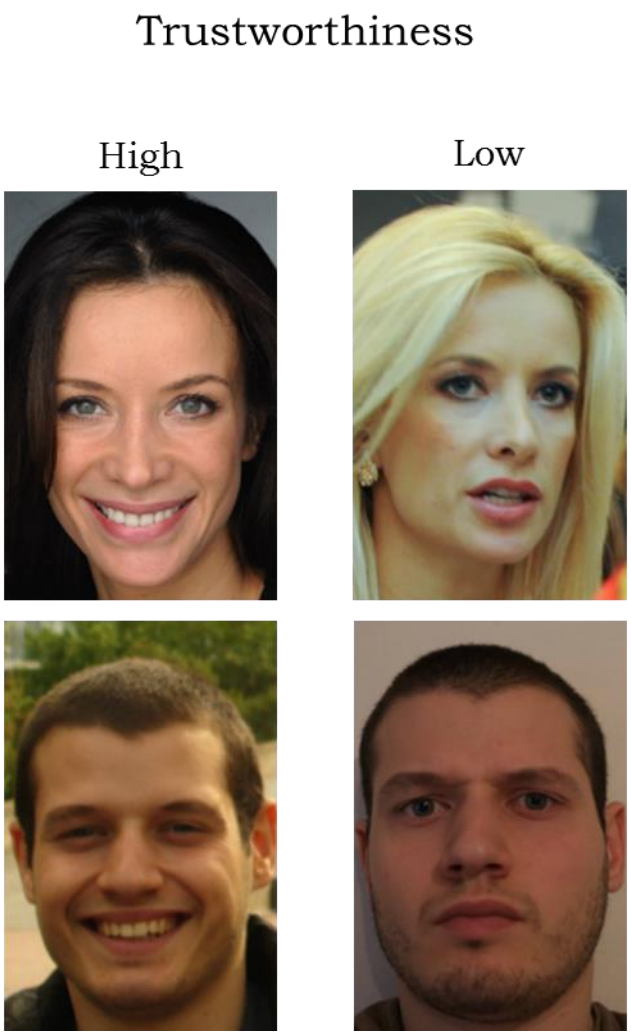

Figure 5.3. Different images of the same people rated high and low in trustworthiness.

Results and discussion

Mean ratings by condition are shown in Table 5.2. A 2x2 within subjects ANOVA revealed significant main effects of face trustworthiness (F (1, 39) = 99.64, p < .001, η_p^2 = .72) and voice trustworthiness (F (1, 39) = 18.03, p < .001, η_p^2 =.32), with no significant interaction (F (1, 39) = 3.19, p > .05, η_p^2 = .08).

Table 5.2. *Mean Ratings of Dominance across Conditions in Experiment 16. SDs in Parentheses.*

	Low trustworthiness voice	High trustworthiness voice
Low trustworthiness face	4.8 (.55)	5.4 (.51)
High trustworthiness face	6.0 (.47)	6.2 (.53)

As with judgements of dominance (Experiment 13), we here show clear, independent contributions of face and voice to multimodal judgements of trustworthiness. However, unlike judgements of dominance, we see in this experiment that faces have the larger effect for trustworthiness attribution. This is consistent with findings from correlational studies which show that the judgement of multimodal stimuli can be influenced more or less by faces and voices, according to the attribute required (Rezlescu et al., 2015).

5.7 General Discussion

In this chapter, we investigated the effect and automaticity of audiovisual integration in social trait attribution. Our results demonstrate that mean vocal pitch is a significant factor for the perception of dominance in voices and that large within-person differences exist in social attribute ratings for faces. Moreover, while both face and voice cues influenced social trait attribution significantly, the relative contribution of the auditory and visual channel to social evaluation was shown to be dependent on the specific social trait. While vocal information was more diagnostic for dominance perception, face information was more diagnostic for the perception of trustworthiness. We also show that audiovisual integration is, to some extent, automatic and that participants cannot completely ignore the audio or visual channel, even when they are instructed to do so.

Results from these experiments reflect findings from previous research which highlight a stronger and more consistent link between mean pitch and

dominance perception than between pitch and trustworthiness perception. Our findings further extend the literature by demonstrating that lowered pitch is associated with perceptions of higher dominance regardless of the gender of speaker. This is consistent with Ohala (1982) as well as some more recent studies (Borkowska & Pawlowski, 2011; Jones et al., 2010; however see McAleer et al., 2014 for different findings). Mean vocal pitch is, therefore, not only an important signal in determining the age, gender, or mood of a speaker (Latinus & Belin, 2011) but it seems that is can also influence the perception of key social attributes such as dominance. Research on pitch and trustworthiness perception is much less consistent, with some studies reporting lower pitch leading to higher ratings of trustworthiness (Tigue et al., 2012), some reporting higher pitch to be perceived as more trustworthy (McAleer et al., 2014) and others failing to find any association between pitch and trustworthiness (Klofstad et al., 2012; Vukovic et al., 2011). Our findings are consistent with the latter group of studies as we did not find a significant association between pitch and trustworthiness. Nevertheless, pitch is one of the many acoustic vocal parameters and our audiovisual integration studies show that vocal information has a significant effect on trustworthiness attribution. This implies there might be other acoustic measures worth exploring such as harmonic-to-noise ratio, which has previously been found to predict ratings of trustworthiness for both male and female speakers (McAleer et al., 2014).

In terms of multimodal social evaluation, our results show clear differences in the relative contribution of auditory and visual cues to social perception for the two fundamental social dimensions – trustworthiness and dominance. Both the face and the voice had a significant effect on trait attribution, however, while audio information was much more diagnostic of dominance perception, the reverse pattern was observed for trustworthiness where facial cues were much more important. Our results on multimodal dominance perception replicate and support the findings of Rezlescu et al. (2015), but oppose studies on the facial overshadowing effect (Tomlin, Stevenage, & Hammond, 2016) which show an advantage for visual information in identity recognition. This highlights the importance of both

context and task demands, and is consistent with face and voice models proposing that identity, affect and speech information is processed along functional pathways which are mostly independent, yet have some scope to interact with one another (Belin et. al, 2011; Young & Bruce, 2011). The importance of auditory information for the perception of dominance and aggression could be due to its higher reliability. Dominance judgements have been shown to correlate highly with sexually dimorphic aspects, and vocal pitch is a sexually dimorphic aspect of voice (Puts et al., 2006). It might, therefore, be a more reliable channel when assessing someone's masculinity, which in turn makes it a salient dominance cue (Collignon, 2008).

Our findings regarding trustworthiness perception, on the other hand, are in contrast to Rezlescu et al. (2015), who found that the facial and vocal channels contributed equally to the perception of trustworthiness and interacted with one another. This might be due to the different facial and especially vocal stimuli used in the present experiments, as we opted to use contentful speech rather than brief neutral vowel sounds. A consistent finding in the face evaluation literature is that social judgements are highly dependent on emotional expressions and that participants often assign a particular emotional expression to seemingly neutral faces (Said et al., 2009). Our findings may therefore indicate that the visual channel is more reliable for extracting emotional content (Massaro & Egen, 1996).

We also show that the combination of auditory and visual cues is mandatory and bidirectional. Such results are consistent with studies of audiovisual integration in emotion and identity recognition (de Gelder & Vroomen, 2000; Schweinberger et al., 2007), all of which imply that combining cross-modal information is not under attentional control. It would appear as though presenting faces and voices together, regardless of task and synchronicity, leads to an automatic integration rather that prompting perceivers to make an explicit decision about whether to integrate the presented information or not. The evidence for the automaticity of audiovisual integration is particularly compelling here, as the voices in the present studies were paired with static faces. While this unquestionably

misrepresents real-life social interactions, it provides a clear indication of the magnitude of this effect – a finding further supported by studies reporting automatic integration even when there was a mismatch in the gender of the face and voice that participants were presented with (Green, Kuhl, Meltzoff, & Stevens, 1991).

Such findings demonstrate clear differences in the weighting of auditory and visual cues in social perception, dependent on the specific social attribute being evaluated. While vocal information is more important for the perception of dominance, facial information has a greater influence on listener attributions of trustworthiness. Furthermore, using a focused-attention paradigm, we show that audiovisual integration appears to be an automatic, bidirectional process. This extends and contributes to the scarce literature on multimodal social evaluation. By using contentful utterances as vocal stimuli, we obtained listener evaluations of speech that represent everyday social interactions more accurately. Moreover, we used images of the same people in both the high and low dominance and trustworthiness conditions and found significant differences between them. This demonstrates that sufficient within-person variability exists in ratings of different images of the same identity, and implies that social evaluation is not only a function of identity but also a function of the properties of images, and so changeable over time. Our social perception of other individuals is flexible and dynamic. As both face and voice models suggest a somewhat independent processing of identity and emotion information in separate pathways, investigating social person evaluation can provide us with essential insight into the possible interaction between those pathways. Combining faces and voices together, therefore, can better inform our knowledge of both audiovisual integration and general models of face and voice processing, alongside as bringing us closer to understanding integrated person perception.

Chapter 6 – Summary and Conclusions

This last chapter brings all findings together and summarises the key contributions of the work included in this thesis. It starts with a brief summary of the results from each experimental chapter, followed by a discussion of the most significant findings. Finally, I highlight the importance of social evaluation as a research area and outline directions for future study.

6.1 Summary of Aims and Results

The overall aim of this thesis was to explore social evaluation and its dependence on within-person variability. Chapter 2 focused on three potential factors of attribution – gender, image averaging, and familiarity. As most first impression studies have been designed to identify the underlying information in the face people use to inform their social judgements, there is not much known about the relationship between social attributes across gender. Results from Chapter 2 showed clear differences for male and female faces, characterised by a negative relationship between attractiveness and dominance as well as between trustworthiness and dominance for female faces. No such pattern was found for male faces. Such results are consistent with social stereotype studies (Sutherland et al., 2015) and together demonstrate that key social attributes, such as dominance, might be assigned a different meaning by perceivers depending on the gender of the target. Findings from Chapter 2 also revealed that the process of image averaging, proposed as a way of providing a more accurate representation of identity, brings about significant changes in social evaluation. Average faces were generally perceived as more attractive and trustworthy (female faces only) as well as less distinctive and dominant (male faces only). Such results support our previous findings on gender differences and can easily be accounted for by the image artefacts associated with averaging, such as blurring and smoothing out of face texture. Finally, despite the fact that familiarity may seem to be an irrelevant factor, it can reveal how evaluation changes as we get to know people better. Results showed that familiarity reduces gender biases and that different images of familiar identities are

rated more similarly than those of unfamiliar identities. This implies that familiarity takes over social evaluation mechanisms and perceivers use their knowledge of and experiences with the target to guide their evaluations rather than the properties of images.

Chapter 3 aims to investigate the spread and magnitude of within- and between-person variability in social evaluation. Experiments in this chapter use a data-driven PCA approach to extract the information in the face diagnostic for different social traits. We firstly incorporate within- and between-person variability together by using 20 different images of 20 identities. Results demonstrated a lot more between-person variability in judgements of attractiveness, but comparable within- and between-person variability in judgements of trustworthiness and dominance. The same procedure was then applied to many images of the same identity in order to investigate whether idiosyncratic variability alone can bring about changes in social evaluation. Using this within-person variability to successfully manipulate the way people are perceived extends previous first impression models by demonstrating that social evaluation is a function of both identity and the statistical properties of images.

Chapter 4 aimed to explore the relationship between social evaluation and identity recognition. It was based on findings that evaluating faces on social dimensions leads to a deeper level of processing and thus improves face memory. Results from this chapter showed only a minimal association between social judgements and perceptual identity tasks such as face matching. Furthermore, experiments in this chapter focused on a common factor between face recognition and social evaluation – emotional expression. They explore the possibility that a smiling expression might reveal additional identity-diagnostic information that aids recognition. Results showed improvements in matching accuracy when both images in the face pair had a smiling, rather than a neutral expression. Critically, an increase was seen in both match and mismatch trials which makes this approach superior to other improvement methods such as training and feedback (Alenezi & Bindemann, 2013; Towler, White, & Kemp, 2017).

Finally, Chapter 5 aimed to extend first impressions from faces to a more integrated person evaluation. Here, we explored the relative contribution of face and voice cues to social evaluation. Experiments in this chapter use within-person variability to manipulate faces and mean vocal pitch to manipulate voices. Together, ratings of these face-voice pairings showed that different cues are critical for different social traits. While vocal information is more important for the perception of dominance, facial information is more critical when it comes to trustworthiness evaluation. Moreover, we show that this audiovisual integration is automatic as participants seem to use both the visual and auditory channels even when instructed to ignore one of them.

6.2 Key Findings

Within-person variability

The main finding of this thesis concerns the role of within-person variability in social evaluation. It is implicated and utilised throughout all experimental chapters, however Chapter 3 specifically demonstrates the spread and magnitude of within-person variability. It suggests that trait inferences depend on the choice of a photograph just as much as they do on the identity represented in that photograph. Experiments in this chapter even showed that it is possible to manipulate social evaluation just by sampling one's idiosyncratic variability. Moreover, experiments in Chapter 4 showed that aspects of this identity-specific variability, such as emotional expressions, could change the face in a way that improves face recognition. Such an approach extends previous evaluation models and allows us to disentangle the effects of identity and image properties on social evaluation, which could not be addressed by merely sampling a single image per identity.

Early theoretical work by Secord in the 1950s (see Secord, 1958 for a review) identified five key inference mechanisms. These address both cultural cues, such as stereotypes relating to age, gender, or race, and expressive cues, such as emotional expressions. The first mechanism, referred to as 'temporal extension', describes the tendency to overgeneralise a momentary

state as reflecting internal enduring personality characteristics. The second, 'parataxis', describes trait inferences based on similarity to or past experience with familiar others. Thus, 'temporal extension' is closely related to emotion overgeneralisation (see Intro), whereas 'parataxis' can be associated with the familiar face overgeneralisation hypothesis. The third mechanism is the only one that does not make use of facial cues, but rather addresses first impressions based on already assigned social categories. Finally, the last two mechanisms, 'functional inference' and 'metaphorical generalisation', link social inferences to facial cues with functional and metaphorical significance. Following from 'functional inference' a person with a larger mouth may be perceived as more social and talkative. 'Metaphorical generalisation', on the other hand, might suggest that people with more redness in their face will be evaluated as more dominant due to the metaphorical link between the colour red and anger and the close relationship between aggressiveness and dominance. All of these five mechanisms have been supported by both classical and more recent studies (Lewicki, 1985; Secord & Jourard, 1956; Verosky & Todorov, 2010a, 2010b, 2013; Young, Elliot, Feltman, & Ambady, 2013; Zebrowitz et al., 2011; Zebrowitz & Montepare, 2008). It is therefore, surprising that current social evaluation models fail to address and account for all of them. Integrating the within-person approach, however, could allow us to explore all key processes and therefore provides a much more accurate representation of reality. Within a single identity we can find images with different emotional expressions, images that remind perceivers of different known identities and even images where the size and shape of certain facial features might look completely different due to facial cosmetics, camera angle, or changes in lighting. Nevertheless, it is important to note that the within-person approach is not an alternative to previous research. Incorporating within-person variability is rather complementary to already existing models. This undoubtedly helps us gain deeper understanding of social face evaluation. Studies incorporating both within- and between-person variability (such as Experiment 5), therefore are in a much better position to achieve a more complete understanding of face perception by ensuring a full range of real-life face variability.

Social evaluation across modality

In an attempt to include even more real life cues, relevant to social evaluation, experiments in Chapter 5 investigated how the visual and auditory channels combine when forming first impressions. Results showed that their effect was additive, rather than interactive. This fits well with face and voice identity models (Belin et. al, 2011; Young & Bruce, 2011) which suggest that identity, affect, and speech information is processed somewhat independently in the brain, with little scope to interact with one another. While both faces and voices contributed to trait inferences significantly, there were clear differences in the weighting of their contribution. Such findings demonstrate that audiovisual integration in social evaluation might have a much more complex mechanism than identity or emotion recognition which highlights the need for further investigation.

Furthermore, experiments in Chapter 5 showed that audiovisual integration is automatic, suggesting that in everyday situations first impressions are probably never based on facial information alone. Models of social evaluation should, therefore, be able to account for both the auditory and the visual channel. This has been already implemented in identity recognition (Belin, Fecteau, & Bedard, 2004) and person construal models (Freeman & Ambady, 2011), however it is not yet addressed in social evaluation.

Ecological validity

What is common across many aspects of this thesis is the use of natural and more realistic stimuli. While in face evaluation this is executed by collecting 'ambient' images that vary in pose, emotional expression, lighting, etc., voice evaluation was explored using contentful, relevant, and meaningful utterances. Like within-person variability, 'ambient' images are not suggested as an alternative to controlled image sets, which could certainly allow a more systematic investigation. Nevertheless, naturalistic images offer a more ecologically valid approach that is more representative of real world social evaluation. Experiments in this thesis demonstrate the utility and importance of natural image variability and suggest that future face perception studies and models should be able to account for it. This

extends the original findings of Jenkins et al. (2011) who argue that natural variability is a critical aspect of face perception, as well as later work by Sutherland et al. (2013) who showed that using 'ambient' images could have significant implications for the mapping of the social evaluation space.

Moreover, there is a stark contrast between the voice stimuli used in Chapter 5 and those used in previous voice evaluation literature. Most such studies have used either long irrelevant passages of speech (Montepare & Zebrowitz-McArthur, 1987; Zuckerman & Driver, 1989) or recordings of people pronouncing vowels in a happy/sad/angry or trustworthy/dominant way (Rezlescu et al., 2015; Tsankova et al., 2015). While this ensures trait inferences based on pure vocal cues, it is not an accurate representation of real life social interactions and first impression situations. Therefore, by using 'ambient' images and contentful voice utterances, experiments in this thesis support the importance of natural variability as a way to study social evaluation using a wide and representative range of cues.

Dimension interpretation

Findings from this thesis are particularly relevant to the mapping out of social evaluation dimensions and the meaning assigned to them by perceivers. Chapter 3 showed clear differences in the spread of ratings for trustworthiness and dominance, on one side, and attractiveness on the other. There was a comparable amount of within- and between-person variability for ratings of trustworthiness and dominance, however ratings of attractiveness varied a lot more between images of different identities than between different images of the same identity. This close link between attractiveness and identity was further supported in Chapter 4 where differences in attractiveness ratings only were associated with face matching performance. Such findings suggest that the dimensions identified by Oosterhof and Todorov (2008) and the one added later by Sutherland et al. (2013) might be relying on different mechanisms.

Oosterhof and Todorov (2008) suggest that trustworthiness is primarily a judgement of a target's intentions and dominance is primarily a judgement

of a target's capability. Together they offer the perceiver an assessment of the target's threat. They argue that the importance of these dimensions has an evolutionary origin associated with the rapid detection of harm. Moreover, similar interpersonal and intergroup models also consist of two fundamental dimensions – warmth and competence which also refer to intentionality and capability (Fiske et al., 2007). Using ambient images to explore social evaluation dimensionality, Surtherland et al. (2013) challenged Oosterhof and Todorov's account. They identified an additional dimension – youthful-attractiveness. While it is easy to imagine that this dimension could also have an evolutionary basis (e.g. sexual selection), it does not seem to relate to the functional evaluation of threat in any way. Therefore, the youthful-attractiveness dimension could be qualitatively different from trustworthiness and dominance.

In addition to the results reported in this thesis, this interpretation is further supported by previous research using fewer and less variable images of the same identity (Sutherland, Young, & Rhodes, 2016; Todorov & Porter, 2014). This seems somewhat surprising and counterintuitive as we generally think of attractiveness as something tied to differences in facial features and image properties, whereas trustworthiness and dominance seem rather linked to identity. Findings from this thesis suggest that a different meaning might be assigned to being, say, a trustworthy person compared to a trustworthy-looking image of that person, whereas these two concepts seem to overlap when it comes to attractiveness.

6.3 Importance and Future Directions

Why is social evaluation important?

The most prevalent criticism of social evaluation is that such judgements are trivial and superficial. Indeed, even Secord himself reported arguments that research on trait inferences provides unsurprising results that can easily be accounted for by common sense (Secord, 1958). In order to counteract this, however, he pointed out that the goal of social science is to explore and quantify significant relationships in the world, rather than

discover new ones. The strength of this line of work comes from perceivers' consistency. The fact that people have been shown to agree with each other's social ratings demonstrate that social evaluation is closely linked to physical changes in the face (or changes in image characteristics as demonstrated in Chapter 3). This then allows the use of data–driven approaches to identify the underlying information in the face people use to inform their judgements.

Another challenge for social evaluation is that evidence for its accuracy is mixed at best (Rule et al., 2013; see Todorov et al., 2015 for a review). This, however, can be overcome by the high levels of inter-rater reliability as well as the great many studies demonstrating the importance of social evaluation. Trait inferences have been shown to predict mate and dating decisions, political outcomes, online financial lending and court decisions, even when people are provided with other information, relevant to their decision. This inevitably makes first impressions not only interesting but also important to fully understand, regardless of their accuracy.

Accuracy

The effect of within-person variability on social evaluation has potential implications for the accuracy of trait judgements. Existing studies present inconsistent results with some reporting small but reliable correlations between first impressions and self-reported personality characteristics (Porter et al., 2008; Rule & Ambady, 2008, 2010), while others report no relationship whatsoever when facial cues such as age, gender and race are taken into consideration (Olivola & Todorov, 2010b). Findings from Todorov and Porter (2014) as well as Chapter 3 demonstrate a great amount of variance in ratings of attractiveness, trustworthiness, and dominance (especially for the latter two) for different images of the same individual. This challenges evidence supporting the accuracy of social judgements, implying that social evaluation might be guided not only by identity, but also by image characteristics and momentary changes in the face. Therefore, such first impressions might be a more accurate representation of situational and momentary intentions, rather than internal personality predispositions. Moreover, the image-dependent nature of social evaluation has been

supported by studies demonstrating that trait inferences represent reality accurately. Verplaetse, Vanneste, and Braeckman (2007), for example, asked participants to play a one-shot prisoner dilemma game to assess their cooperativeness and used a webcam to take a picture of them at the exact moment they made their decision. These images, together with a new set of photographs of the same identities taken prior to participation, were then rated by unfamiliar others. Results showed that it was possible to accurately discriminate cooperative and non-cooperative players but only by using the images taken at the decision-making moment.

While challenging existing evidence for the accuracy of social judgements, within-person variability can also provide a possible mechanism to improve the correspondence between trait inferences and stable personality characteristics. As within-person social evaluation involves collecting ratings of many different images of the same identity, it can easily be related to swarm intelligence (Krause, Ruxton, & Krause, 2010) or the wisdom of crowds (Budescu & Chen, 2014; Davis-Stober, Budescu, Dana, & Broomell, 2014). These phenomena are based on the earlier work of Galton (1907) demonstrating that the average estimate made by a group of people is usually very close to the veridical. Such an approach is particularly well suited for more difficult tasks, characterised by large variation in responses (Krause, James, Faria, Ruxton, & Krause, 2011). Kerr and Tindale (2004) even showed that the average estimate of the group can be more accurate than the estimate of the best performers. The wisdom of crowds can also be applied to face recognition, where individual performance on a face matching task can be substantially improved by aggregating the data from groups of eight and above (White, Burton, Kemp, & Jenkins, 2013). In order to address trait inferences from faces, we just need to substitute groups of participants with groups of images depicting the same individual. Averaging across ratings of these images then, may be a more accurate representation of reality and reveal more about the individual and his or her personality. Using such an approach can help perceivers detect general tendencies and predisposition patterns that might be more informative of stable personality traits. Thus, a person who is smiling in most of their images might be more likely to be

genuinely friendlier and more approachable. This within-person approach simulates real life situations where we refine our idea of someone's character every time we encounter them and interact with them.

Own social evaluation

With the high levels of inter-rater agreement in social judgements, it can be argued that perceivers use particular features, combinations of features or even image properties to make such decisions. It is, therefore, interesting to establish whether these mechanisms can be applied to images of the perceivers themselves, i.e. can participants accurately detect images that will be perceived as more attractive, trustworthy, or dominant by others? Findings from Chapter 2 challenge this suggestion as they show that images of familiar identities are rated much more similarly compared to images of unfamiliar identities. This suggests that familiarity might make us blind to factors affecting social evaluation, hence making us unable to choose images that lead to the desired social perception. This is supported by identity studies asking participants to select veridical and manipulated images of themselves that are their best representation. Such studies report a tendency for participants to select artificially enhanced images of their own face (Allen, Brady, & Tredoux, 2009) as well as of other familiar identities (Lee & Perrett, 2000). People seem to be unable to select their own best-likeness image even when they are not artificially modified in any way. White, Burton, and Kemp (2015), for example, asked participants to rate their own images for best likeness and then collected the same ratings from another sample of participants who were briefly familiarised with the target identities. Using these images in a matching task revealed divergent perceptions of likeness such that images selected by familiarised others led to higher matching accuracy than images chosen by the targets themselves.

Moreover, studies combining both identity and social evaluation also show that participants tend to choose images manipulated to look more attractive and trustworthy as their best likeness (Epley & Whitchurch, 2008; Verosky & Todorov, 2010b; Zell & Balcetis, 2012). This is consistent with studies showing that participants intentionally try to select images for online

social networks that will be perceived more favourably by others (Siibak, 2009) and that self-selected dating profile images are evaluated as more attractive than images taken under controlled lab conditions (Hancock & Toma, 2009). While this demonstrates people's understanding of first impressions and their importance, evidence from the identity literature and Chapter 2 indicate that we might be less sensitive to the factors affecting social evaluation when it comes to images of ourselves or familiar others. Indeed, evidence from a recent study by White, Sutherland, and Burton (2017) showed that self-selected profile images for social, dating, and professional websites were evaluated less favourably than images selected by strangers.

Integrated person evaluation

The main aim of this thesis was to explore the role of natural variability in social evaluation. While this was addressed by sampling naturalistic images and contentful and meaningful voice recordings, covering both between- and within-person variability, there are additional person cues that might be of interest to future research. Experiments in Chapter 5 focus on a specific vocal characteristic – mean pitch. It was found to have a great effect on the attribution of dominance, but not on trustworthiness evaluation. Interestingly, auditory cues still had a significant effect on ratings of trustworthiness, implying there might be other vocal qualities relevant to trustworthiness (Hodges-Simeon et al., 2010). McAleer et al. (2014), for example, not only showed that first impressions from voices follow the same two-dimensional structure as first impressions from faces, but they also explored possible acoustic properties related to those dimensions. Their findings showed that mean pitch was a significant predictor of dominance for both male and female voices, but it predicted trustworthiness ratings in male voices only. Harmonic-to-noise ratio (a measure of the hoarseness of the voice), on the other hand, was a significant predictor of trustworthiness for both male and female voices making it the most reliable voice cue for this trait. Other vocal properties implicated in trustworthiness evaluation included voice glide (related to the constriction of airflow through the vocal tract) and intonation, whereas formant dispersion (an acoustic correlate of

vocal tract length) and alpha ratio (measure of the source spectral slope) were additional significant factors of dominance. Moreover, accent has been related to both categorisation and social perception (Hansen, Rakic, & Steffens, 2017; Rakic, Steffens, & Mummendey, 2011), which makes it another interesting cue for future research. Further investigation of all these acoustic measures will certainly tell us more about the mechanisms behind first impressions from voices and will reveal another layer of social person evaluation.

Apart from facial and vocal properties, body cues have also been shown to influence person evaluation. Studies have demonstrated that people use body cues to recognise others (Rice, Phillips, Natu, An, & O'Toole, 2013; Rice, Phillips, & O'Toole, 2013), to assess their health and emotional state (Aviezer, Trope, & Todorov, 2012; de Gelder, de Borst, & Watson, 2015; Puhl & Heuer, 2009), and to evaluate their attractiveness and potential as a mate (Currie & Little, 2009; Peters, Rhodes, & Simmons, 2008). Furthermore, Hill et al. (2016) used PCA on a range of body descriptors sourced from online dating websites and clothing retailer fit recommendations to establish five key dimensions of body evaluation – weight, height, femininity, masculinity, and waist height (see Figure 6.1 for dimension examples).

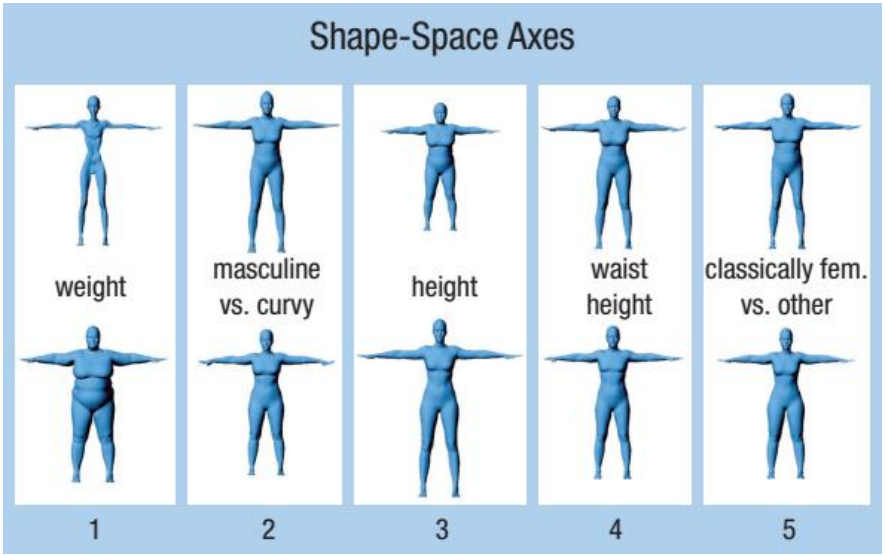

Figure 6.1. The first five dimensions from the body shape space in Hill et al. (2016). For each component the body on the top is 3 SDs above the original and the body on the bottom is 3 SDs below the original.

Linking this body cue space to first impression dimensions would be a natural next step in this line of work, bringing us even closer to understanding integrated person evaluation. It is worth noting that this has already been considered in attractiveness ratings (Saxton, Burriss, Murray, Rowland, & Roberts, 2009), however, the integration of face, voice, and body cues has not been explored in dominance or trustworthiness attribution.

6.4 Overall Conclusions

In summary, this thesis aimed to explore social evaluation across gender, familiarity and modality as well as compare the influence of between- and within-person variability. Results showed that different images of the same person can vary just as much as images of different identities when it comes to ratings of trustworthiness and dominance. Moreover, I showed that it is possible to manipulate the way someone is socially perceived by sampling their own idiosyncratic variability. This demonstrates that social evaluation depends on both identity and image properties, highlighting how important it is for future social evaluation models to address both sources of variability.

The work described here also integrates different modalities and adopts a more naturalistic approach by using 'ambient' images and contentful voice utterances. Findings revealed clear differences in the weighting of face and voice cues in the evaluation of the two fundamental dimensions – trustworthiness and dominance. While visual information from the face is more diagnostic for trustworthiness evaluation, ratings of dominance seem to be guided by auditory information to a greater extent. This adds another layer of complexity to already existing first impression models and brings this line of research a step closer to understanding integrated person evaluation.

References

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology, 93*(5), 751–763. doi: 10.1037/0022-3514.93.5.751

Ahumada, A. J., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America, 49(*6B), 1751–1756. http://dx.doi.org/10.1121/1.1912577

Aishwarya, P., & Marcus, K. (2010). Face recognition using multiple eigenface subspaces. *Journal of Engineering and Technology Research, 2*(8), 139–143.

Albright, L., Malloy, T. E., Dong, Q., Kenny, D., Fang, X., Winquist, L., & Yu, D. (1997). Cross-cultural consensus in personality judgments. *Journal of Personality and Social Psychology*, *72*(3), 558–569. http://dx.doi.org/10.1037/0022-3514.72.3.558

Alenezi, H. M., & Bindemann, M. (2013). The effect of feedback on face-matching accuracy. *Applied Cognitive Psychology, 27*(6), 735–753. doi: 10.1002/acp.2968

Allen, H., Brady, N., & Tredoux, C. (2009). Perception of 'best likeness' to highly familiar faces of self and friend. *Perception, 38*(12), 1821–1830. doi: 10.1068/p6424

Allport, G. W., & Cantril, H. (1934). Judging personality from voice. *The Journal of Social Psychology, 5*(1), 37–55. http://dx.doi.org/10.1080/00224545.1934.9921582

Apicella, C. L., Feinberg, D. R., & Marlowe, F. W. (2007). Voice pitch predicts reproductive success in male hunter-gatherers. *Biology Letters*, *3(*6), 682–684. doi: 10.1098/rsbl.2007.0410

Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, *37*(5), 715–727. http://dx.doi.org/10.1037/0022-3514.37.5.715

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, *338*(6111), 1225–1229. doi: 10.1126/science.1224313

Bailenson, J. N., Iyengar, S., Yee, N., & Collins, N. A. (2009). Facial similarity between voters and candidates causes influence. *Public Opinion Quarterly*, *72*(5), 935–961. https://doi.org/10.1093/poq/nfn064

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

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion, 6*(2), 269–278. http://dx.doi.org/10.1037/1528-3542.6.2.269

Baudouin, J. Y., & Humphreys, G. W. (2006). Configural information in gender categorisation. *Perception*, *35*(4), 531–540. doi: 10.1068/p3403

Bayliss, A. P., & Tipper, S. P. (2006). Predictive gaze cues and personality judgments should eye trust you? *Psychological Science*, *17*(6), 514–520. doi: 10.1111/j.1467-9280.2006.01737.x

Beard, B. L., & Ahumada Jr, A. J. (1998). Technique to extract relevant image features for visual tasks. In *Photonics West'98 Electronic Imaging* (pp. 79–85). International Society for Optics and Photonics.

Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology, 102*(4), 711–725. doi: 10.1111/j.2044-8295.2011.02041.x

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

Benson, P. J., & Perrett, D. I. (1993). Extracting prototypical facial images from exemplars. *Perception*, *22*(3), 257–262. doi: 10.1068/p220257

Berry, D. S. (1990). Vocal Attractiveness and vocal babyishness - effects on stranger, self, and friend impressions. *Journal of Nonverbal Behavior, 14*(3), 141–153. doi: 10.1007/BF00996223

Berry, D. S. (1991). Accuracy in social perception: contributions of facial and vocal information. *Journal of Personality and Social Psychology*, *61*(2), 298–307. http://dx.doi.org/10.1037/0022-3514.61.2.298

Berry, D. S., & McArthur, L. Z. (1986). Perceiving character in faces: The impact of age-related craniofacial changes on social perception. *Psychological Bulletin*, *100*(1), 3–18. http://dx.doi.org/10.1037/0033-2909.100.1.3

Beveridge, J. R., Givens, G. H., Phillips, P. J., & Draper, B. A. (2009). Factors that influence algorithm performance in the face recognition grand challenge. *Computer Vision and Image Understanding*, *113*(6), 750–762. https://doi.org/10.1016/j.cviu.2008.12.007

Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The effect of image pixelation on unfamiliar-face matching. *Applied Cognitive Psychology, 27*(6), 707–717. doi: 10.1002/acp.2970

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science, 15*(10), 674–679. doi: 10.1111/j.0956-7976.2004.00739.x

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, New York: Addison-Wesley/ACM Press, pp. 187–194. doi: 10.1145/311535.311556

Boersma, P., & Weenink, D. (2016). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.22.

Bonner, L., Burton, A. M., & Bruce, V. (2003). Getting to know you: How we learn new faces. *Visual Cognition*, *10*(5), 527–536. http://dx.doi.org/10.1080/13506280244000168

Boothroyd, L. G., Jones, B. C., Burt, D. M., & Perrett, D. I. (2007). Partner characteristics associated with masculinity, health and maturity in male faces. *Personality and Individual Differences*, *43*(5), 1161–1173. https://doi.org/10.1016/j.paid.2007.03.008

Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour, 82*(1), 55–59. http://dx.doi.org/10.1016/j.anbehav.2011.03.024

Bower, G. H., & Karlin, M. B. (1974). Depth of processing pictures of faces and recognition memory. *Journal of Experimental Psychology*, *103*(4), 751–757. http://dx.doi.org/10.1037/h0037190

Brahnam, S. (2005). A computational model of the trait impressions of the face for agent perception and face synthesis. *AISB Journal*, *1*(6), 481–508.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436. http://dx.doi.org/10.1163/156856897X00357

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*(1), 105–116. doi: 10.1111/j.2044-8295.1982.tb01795.x

Bruce, V. (1986). Influences of familiarity on the processing of faces. *Perception*, *15*(4), 387–397. doi: 10.1068/p150387

Bruce, V., Burton, A. M., & Dench, N. (1994). What's distinctive about a distinctive face? *Quarterly Journal of Experimental Psychology*, *47*(1), 119–141. http://dx.doi.org/10.1080/14640749408401146

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*(4), 339–360. http://dx.doi.org/10.1037/1076-898X.5.4.339

Bruce, V., Henderson, Z., Newman, C., & Burton. A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7*(3), 207–218. doi: 10.1037/1076-898X.7.3.207

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

Bruckert, L., Liénard, J. S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice parameters to assess men's characteristics. *Proceedings of the Royal Society of London B: Biological Sciences*, *273*(1582), 83–89. doi: 10.1098/rspb.2005.3265

Buckingham, G., DeBruine, L. M., Little, A. C., Welling, L. L., Conway, C. A., Tiddeman, B. P., & Jones, B. C. (2006). Visual adaptation to masculine and feminine faces influences generalized preferences and perceptions of trustworthiness. *Evolution and Human Behavior*, *27*(5), 381–389. https://doi.org/10.1016/j.evolhumbehav.2006.03.001

Budescu, D. V., & Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280.

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. doi: 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. doi: 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*(3), 256–284. https://doi.org/10.1016/j.cogpsych.2005.06.003

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology, 102*(4), 943–958. doi: 10.1111/j.2044-8295.2011.02039.x

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*(1), 202–223. doi: 10.1111/cogs.12231

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291. doi: 10.3758/BRM.42.1.286

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*(3), 243–248. doi: 10.1111/1467-9280.00144

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, *41*(9), 1179–1208. doi: 10.1016/S0042-6989(01)00002-5

Calder, A. J., Young, A. W., Keane, J., & Dean, M. (2000). Configural information in facial expression perception. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 527–551. http://dx.doi.org/10.1037/0096-1523.26.2.527

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(2), 535–543. http://dx.doi.org/10.1016/j.tics.2007.10.001

Carre, J. M., & McCormick, C. M. (2008). In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society B, 275*(1651), 2651–2656. doi: 10.1098/rspb.2008.0873

Carre, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science, 20*(10), 1194–1198. doi: 10.1111/j.1467-9280.2009.02423.x

Changizi, M., Zhang, Q., & Shimojo, S. (2006). Bare skin, blood and the evolution of primate colour vision. *Biology Letters, 2*(2), 217–221. doi: 10.1098/rsbl.2006.0440

Chen, D., Halberstam, Y., & Yu, A. (2016). Perceived Masculinity Predicts US Supreme Court Outcomes. *PloS One*, *11*(10), e0164324. https://doi.org/10.1371/journal.pone.0164324

Chen, W., Lander, K., & Liu, C. H. (2011). Matching faces with emotional expressions. *Frontiers in Psychology*, *2,* 1–10. https://doi.org/10.3389/fpsyg.2011.00206

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., & Maneewongvatana, S. (2009). Encoding emotions in speech with the size code. *Phonetica*, *65*(4), 210–230. doi: 10.1159/000192793

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, *31*(8), 985–994. doi: 10.1068/p3335

Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition*, *11*(7), 857–869. http://dx.doi.org/10.1080/13506280444000021

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology, 17*(1), 97–116. http://dx.doi.org/10.1080/09541440340000439

Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M.R. (2014). Inferring character from faces: A developmental study. *Psychological Sciences, 25*(5), 1132–1139. doi: 10.1177/0956797614523297

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research, 1242,* 126–135. http://dx.doi.org/10.1016/j.brainres.2008.04.023

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour*, *65*(5), 997–1004. http://dx.doi.org/10.1006/anbe.2003.2123

Courtois, M. R., & Mueller, J. H. (1979). Processing multiple physical features in facial recognition. *Bulletin of the Psychonomic Society*, *14*(1), 74–76. doi: 10.3758/BF03329404

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–45.

Craw, I. (1995). A manifold model of face and object recognition. In T. Valentine (Ed.), *Cognitive and computational aspects of face recognition* (pp. 183–203). London: Routledge.

Crete, F., Dolmiere, T., Ladret, P., & Nicolas, M. (2007). The blur effect: perception and estimation with a new no-reference perceptual blur

metric. In *Electronic Imaging 2007* (pp. 64920I-64920I). International Society for Optics and Photonics.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61–149. doi: 10.1016/S0065-2601(07)00002-0

Cunningham, M. R. (1986). Measuring the physical in physical attractiveness: Quasi-experiments on the sociobiology of female facial beauty. *Journal of Personality and Social Psychology*, *50*(5), 925–935. http://dx.doi.org/10.1037/0022-3514.50.5.925

Cunningham, M. R., Barbee, A. P., & Pike, C. L. (1990). What do women want? Facialmetric assessment of multiple motives in the perception of male facial physical attractiveness. *Journal of Personality and Social Psychology*, *59*(1), 61–72. http://dx.doi.org/10.1037/0022-3514.59.1.61

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.

Currie, T. E., & Little, A. C. (2009). The relative importance of the face and body in judgments of human physical attractiveness. *Evolution & Human Behavior*, *30*(6), 409–416. doi:10.1016/j.evolhumbehav.2009.06.005

D'Argembeau, A., Van der Linden, M., Comblain, C., & Etienne, A. M. (2003). The effects of happy and angry expressions on identity and expression memory for unfamiliar faces. *Cognition & Emotion*, *17*(4), 609–622. http://dx.doi.org/10.1080/02699930302303

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79–101. http://dx.doi.org/10.1037/dec0000004

de Gelder, B., de Borst, A. W., & Watson, R. (2015). The perception of emotion in body expressions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 149–158. doi: 10.1002/wcs.1335

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion*, *14*(3), 289-311. http://dx.doi.org/10.1080/026999300378824

DeBruine, L. M. (2005). Trustworthy but not lust-worthy: Context-specific effects of facial resemblance. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*(1566), 919–922. doi: 10.1098/rspb.2004.3003

DeBruine, L. M., Jones, B. C., Unger, L., Little, A. C., & Feinberg, D. R. (2007). Dissociating averageness and attractiveness: Attractive faces are not always average. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1420–1430. http://dx.doi.org/10.1037/0096-1523.33.6.1420

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 248-255). IEEE.

Dimos, K., Dick, L., & Dellwo, V. (2015). Perception of levels of emotion in speech prosody. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: The University of Glasgow.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*(3), 285–290. http://dx.doi.org/10.1037/h0033731

Doll, L. M., Hill, A. K., Rotella, M. A., Cárdenas, R. A., Welling, L. L., Wheatley, J. R., & Puts, D. A. (2014). How well do men's faces and voices index mate quality and dominance? *Human Nature*, *25*(2), 200–212. doi: 10.1007/s12110-014-9194-3

Dolzycka, D., Herzmann, G., Sommer, W., & Wilhelm, O. (2014). Can Training Enhance Face Cognition Abilities in Middle-Aged Adults? *PLoS ONE*, *9*(3), e90249. https://doi.org/10.1371/journal.pone.0090249

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*(5), 562–571. doi: 10.1177/1948550611430272

Dotsch, R., Wigboldus, D. H. J., Langner, O., & Van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science, 19(*10), 978–980. doi: 10.1111/j.1467-9280.2008.02186.x

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, *106*(3), 433–445. doi: 10.1111/bjop.12103

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

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

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*(1), 109–128. http://dx.doi.org/10.1037/0033-2909.110.1.109

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*(5), 383–386. doi: 10.1111/j.1467-9280.2006.01716.x

Eibl-Eibesfeldt, I. (1989). *Human ethology*. New York, NY: Aldine de Gruyter.

Ekman, P. (1992). Facial expressions of emotion: An old controversy and new findings. *Philosophical Transactions of the Royal Society of London: B. Biological Sciences, 335*(1273), 63–69. doi: 10.1098/rstb.1992.0008

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, *1*(1), 56–75. doi: 10.1007/BF01115465

Ellis, H. D. (1981). Theoretical aspects of face recognition. In Davies et al. (Eds), *Perceiving and remembering faces* (pp. 171–197). London: Academic Press.

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra-and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology*, *88*(1), 143–156. doi: 10.1111/j.2044-8295.1997.tb02625.x

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception*, *8*(4), 431–439. doi: 10.1068/p080431

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience, 19*(9), 1508–1519. doi:10.1162/jocn.2007.19.9.1508

Enlow, D. H., & Hans, M. G. (1996). *Essentials of facial growth*. Philadelphia, PA: WB Saunders Company.

Epley, N., & Whitchurch, E. (2008). Mirror, mirror on the wall: Enhancement in self-recognition. *Personality and Social Psychology Bulletin, 34*(9), 1159–1170. doi: 10.1177/0146167208318601

Fecher, N. (2015). *Praat pitch alteration script*. Department of Language and Linguistics, University of York. Script for Praat.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour*, *69*(3), 561–568. http://dx.doi.org/10.1016/j.anbehav.2004.06.012

Feinberg, D. R., Jones, B. C., Smith, M. L., Moore, F. R., DeBruine, L. M., Cornwell, R. E., … & Perrett, D. I. (2006). Menstrual cycle, trait estrogen level, and masculinity preferences in the human

voice. *Hormones and Behavior*, *49*(2), 215–222. http://dx.doi.org/10.1016/j.yhbeh.2005.07.004

Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin*, *111*(2), 304–341. http://dx.doi.org/10.1037/0033-2909.111.2.304

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

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

Ford, A., & Roberts, A. (1998). Colour space conversions. *Westminster University, London*, *1998*, 1-31.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–279. http://dx.doi.org/10.1037/a0022327

Fruhen, L., Watkins, C. & Jones, B. (2015). Perceptions of facial dominance, trustworthiness and attractiveness predict managerial pay awards in experimental tasks. *The Leadership Quarterly, 26*(6), 1005–1016. http://dx.doi.org/10.1016/j.leaqua.2015.07.001

Galli, G., Feurra, M., & Viggiano, M. P. (2006). "Did you see him in the newspaper?" Electrophysiological correlates of context and valence in face processing. *Brain Research*, *1119*(1), 190–202. https://doi.org/10.1016/j.brainres.2006.08.076

Galton, F. (1879). Composite portraits, made by combining those of many different persons into a single resultant figure. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *8*, 132–144.

Galton, F. (1883). *Inquiries into human faculty and its development*. London: Macmillan.

Galton, F. (1907). Vox populi (The wisdom of crowds). *Nature, 75*(7), 450–451.

Germine, L., Nakayama, K., Duchaine, B., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*(5), 847–857. doi: 10.3758/s13423-012-0296-9

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*(1), 8–20. doi: 10.3758/BF03198438

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experiment Psychology: Learning, Memory, & Cognition*, *16*(1), 5–16. http://dx.doi.org/10.1037/0278-7393.16.1.5

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*(3), 546–567. http://dx.doi.org/10.1037/0033-295X.100.3.546

Gómez-Valdés, J., Hünemeier, T., Quinto-Sánchez, M., Paschetta, C., de Azevedo, S., González, M. F., ... & Bau, C. H. (2013). Lack of support for the association between facial shape and aggression: a reappraisal based on a worldwide population genetics perspective. *PloS One*, *8*(1), e52317. https://doi.org/10.1371/journal.pone.0052317

Gonzalez, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics*, *32*(2), 277–287. http://dx.doi.org/10.1016/S0095-4470(03)00049-4

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261–2271. doi: 10.1016/S0042-6989(01)00097-9

Gosselin, F., & Schyns, P. G. (2004). No troubles with bubbles: A reply to Murray and Gold. *Vision Research*, *44*(5), 471–477. doi: 10.1016/j.visres.2003.10.007

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika, 40*(1), 33–51. doi: 10.1007/BF02291478

Graham, D. J., & Field, D. J. (2007). Statistical regularities of art images and natural scenes: Spectra, sparseness and nonlinearities. *Spatial Vision*, *21*(1), 149–164. PMID: 18073056

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

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics, 50*(6), 524–536. http://dx.doi.org/10.3758/BF03207536

Griffin, A. M., & Langlois, J. H. (2006). Stereotype directionality and attractiveness stereotyping: Is beauty good or is ugly bad? *Social Cognition*, *24*(2), 187–206. doi: 10.1521/soco.2006.24.2.187

Hancock, J. T., & Toma, C. L. (2009). Putting your best face forward: The accuracy of online dating photographs. *Journal of Communication, 59*(2), 367–386. doi: 10.1111/j.1460-2466.2009.01420.x

Hancock, P. J. B., Bruce, V., & Burton, A. M. (1998). A comparison of two computer based face identification systems with human perceptions of faces. *Vision Research, 38*(15), 2277–2288. https://doi.org/10.1016/S0042-6989(97)00439-2

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

Hancock, P. J. B., Burton, A. M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory & Cognition*, *24*(1), 26–40. doi: 10.3758/BF03197270

Hansen, K., Rakić, T., & Steffens, M. C. (2017). Competent and Warm? *Experimental Psychology, 64,* 27–36. http://dx.doi.org/10.1027/1618-3169/a000348

Harries, M. L. L., Walker, J. M., Williams, D. M., Hawkins, S., & Hughes, I. A. (1997). Changes in the male voice at puberty. *Archives of Disease in Childhood, 77*(5), 445–447. http://dx.doi.org/10.1136/adc.77.5.445

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. http://dx.doi.org/10.1016/S1364-6613(00)01482-0

Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues, 57*(4), 657–674. doi: 10.1111/0022-4537.00234

Hess, U., Blairy, S., & Kleck, R. E. (2000). The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal Behavior*, *24*(4), 265–283. doi: 10.1023/A:1006623213355

Hess, U., Kappas, A., & Scherer, K. (1988). Multichannel communication of emotion: Synthetic signal production. In Scherer, K. (Ed.), *Facets of emotion: Recent research* (pp. 161–182). Hillsdale, NJ: Erlbaum.

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception & Performance*, *22*(4), 986–1004. doi: 10.1037/0096-1523.22.4.986

Hill, M. Q., Streuber, S., Hahn, C. A., Black, M. J., & O'Toole, A. J. (2016). Creating body shapes from verbal descriptions by linking similarity spaces. *Psychological Science*, *27*(11), 1486–1497. doi: 10.1177/0956797616663878

Hodges-Simeon, C. R., Gaulin, S. J., & Puts, D. A. (2010). Different vocal parameters predict perceptions of dominance and attractiveness. *Human Nature*, *21*(4), 406–427. doi: 10.1007/s12110-010-9101-5

Hodges-Simeon, C. R., Gaulin, S. J., & Puts, D. A. (2011). Voice correlates of mating success in men: Examining "contests" versus "mate choice" modes of sexual selection. *Archives of Sexual Behavior*, *40*(3), 551–557. doi: 10.1007/s10508-010-9625-0

Hollien, H., Green, R., & Massey, K. (1994). Longitudinal research on adolescent voice change in males. *Journal of the Acoustical Society of America, 96*(5), 2646–2654. http://dx.doi.org/10.1121/1.411275

Hollingworth, H. L. (1922). *Judging human character*. New York: D Appleton.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14*(3), 399–425. doi: 10.1007/s10683-011-9273-9

Hudson, T., De Jong, G., McDougall, K., Harrison, P., & Nolan, F. (2007). F0 statistics for 100 young male speakers of Standard Southern British English. *In Proceedings of the 16th International Congress of Phonetic Science*, Saarbrücken: Germany, 1809-1812.

Humphreys, G. W., Donnelly, N., & Riddoch, M. J. (1993). Expression is computed separately from facial identity, and it is computed separately for moving and static faces: Neuropsychological evidence. *Neuropsychologia*, *31*(2), 173–181. https://doi.org/10.1016/0028-3932(93)90045-2

Hunter, M. D., Phang, S. Y., Lee, K. H., & Woodruff, P. W. (2005). Gender-specific sensitivity to low frequencies in male speech. *Neuroscience Letters*, *375*(3), 148–150. http://dx.doi.org/10.1016/j.neulet.2004.11.003

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*(3), 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*(5862), 435. doi: 10.1126/science.1149656

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society, B., 366*(1571), 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*(3), 313–323. doi: 10.1016/j.cognition.2011.08.001

Joassin, F., Maurage, P., Bruyer, R., Crommelinck, M., & Campanella, S. (2004). When audition alters vision: an event-related potential study of

the cross-modal interactions between faces and voices. *Neuroscience Letters*, *369*(2), 132–137. http://dx.doi.org/10.1016/j.neulet.2004.07.067

Jones, A. L., Russell, R., & Ward, R. (2015). Cosmetics alter biologically based factors of beauty: Evidence from facial contrast. *Evolutionary Psychology, 13*(1), 210–229. doi: 10.1177/147470491501300113

Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C., & Vukovic, J. (2008). Integrating cues of social interest and voice pitch in men's preferences for women's voices. *Biology Letters*, *4*(2), 192–194. doi: 10.1098/rsbl.2007.0626

Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C., & Vukovic, J. (2010). A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour*, *79*(1), 57–62. http://dx.doi.org/10.1016/j.anbehav.2009.10.003

Kaiser, S. B. (1985). *Social psychology of clothing and personal adornment*. Macmillan; Collier Macmillan.

Karremans, J. C., Dotsch, R., & Corneille, O. (2011). Romantic relationship status biases memory of faces of attractive opposite-sex others: Evidence from a reverse-correlation paradigm. *Cognition*, *121*(3), 422–426. https://doi.org/10.1016/j.cognition.2011.07.008

Kaufmann, J. M., & Schweinberger, S. R. (2004). Expression influences the recognition of familiar faces. *Perception*, *33*(4), 399–408. doi: 10.1068/p5083

Keating, C. F. (1985). Gender and the physiognomy of dominance and attractiveness. *Social Psychology Quarterly*, *48*(1), 61–70. http://dx.doi.org/10.2307/3033782

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55*, 623–655. doi: 10.1146/annurev.psych.55.090902.142009

Kirby, M., & Sirovich, L. (1990). Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, *12*(1), 103–108. doi: 10.1109/34.41390

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36*, ECVP Abstract Supplement.

Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society of London B: Biological Sciences*, *279*(1738), 2698–2704. doi: 10.1098/rspb.2012.0311

Knowles, K. K., & Little, A. C. (2016). Vocal fundamental and formant frequencies affect perceptions of speaker cooperativeness. *The Quarterly Journal of Experimental Psychology*, *69*(9), 1657–1675. http://dx.doi.org/10.1080/17470218.2015.1091484

Koch, M., Denzler, J., & Redies, C. (2010). 1/f 2 Characteristics and isotropy in the fourier power spectra of visual art, cartoons, comics, mangas, and different categories of photographs. *PLoS one*, *5*(8), e12268. https://doi.org/10.1371/journal.pone.0012268

Kramer, R. S. (2016). Within-person variability in men's facial width-to-height ratio. *PeerJ*, *4*, e1801.

Kramer, R. S., Jenkins, R., & Burton, A. M. (2016). InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior Research Methods*.

Kramer, R. 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*(4), 1–10. doi: 10.1167/15.4.1

Kramer, S., Zebrowitz, L. A., San Giovanni, J. P., & Sherak, B. (1995). Infant preferences for attractiveness and babyfaceness. *Studies in Perception and Action III*, 389–392.

Kraus, M. W., & Chen, S. (2010). Facial-feature resemblance elicits the transference effect. *Psychological Science*, *21*(4), 518–522. doi: 10.1177/0956797610364949

Krause, J., Ruxton, G. D., & Krause, S. (2010). Swarm intelligence in animals and humans. *Trends in Ecology & Evolution*, *25*(1), 28–34. doi: https://doi.org/10.1016/j.tree.2009.06.016

Krause, S., James, R., Faria, J. J., Ruxton, G. D., & Krause, J. (2011). Swarm intelligence in humans: Diversity can trump ability. *Animal Behaviour, 81*(5), 941–948. https://doi.org/10.1016/j.anbehav.2010.12.018

Krumhuber, E., Manstead, A. S., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, *7*(4), 730–735. http://dx.doi.org/10.1037/1528-3542.7.4.730

Kukkonen, H., Rovamo, J., Tiippana, K., & Näsänen, R. (1993). Michelson contrast, RMS contrast and energy of various spatial stimuli at threshold. *Vision Research*, *33*(10), 1431–1436. https://doi.org/10.1016/0042-6989(93)90049-3

Künzel, H. J. (1989). How well does average fundamental frequency correlate with speaker height and weight? *Phonetica, 46*(1–3), 117–125. doi: 10.1159/000261832

Lakens, D., Fockenberg, D. A., Lemmens, K. P., Ham, J., & Midden, C. J. (2013). Brightness differences influence the evaluation of affective pictures. *Cognition & emotion*, *27*(7), 1225–1246. http://dx.doi.org/10.1080/02699931.2013.781501

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

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, *1*(2), 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*, *5*(4), 214–220. doi: 10.1111/j.1467-9280.1994.tb00503.x

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology, 21*(4), R143–R145. http://dx.doi.org/10.1016/j.cub.2010.12.033

Laver, J. (1994). *Principles of phonetics*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139166621

Lee, J. L., & Perrett, D. I. (2000). Manipulation of colour and shape information and its consequence upon recognition and best-likeness judgments. *Perception, 29*(11), 1291–1312. doi: 10.1068/p2792

Lewicki, P. (1985). Nonconscious biasing effects of single instances on subsequent judgments. *Journal of Personality and Social Psychology*, *48*(3), 563–574. http://dx.doi.org/10.1037/0022-3514.48.3.563

Lindh, J. (2006). Preliminary F0 statistics and forensic phonetics. *Proceedings of the 15th annual International Association of Forensic Phonetics and Acoustics conference*, Department of Linguistics, Göteborg University: Sweden.

Litterer, O. F. (1933). Stereotypes. *Journal of Social Psychology, 4*(1), 59–69. http://dx.doi.org/10.1080/00224545.1933.9921557

Little, A. C., Burt, D. M., & Perrett, D. I. (2006). What is good is beautiful: Face preference reflects desired personality. *Personality and Individual Differences, 41*(6), 1107–1118. http://dx.doi.org/10.1016/j.paid.2006.04.015

Little, A. C., & Hancock, P. J. B. (2002). The role of masculinity and distinctiveness in judgments of human male facial attractiveness. *British Journal of Psychology*, *93*(4), 451–464. doi: 10.1348/000712602761381349

Luevano, V. X., & Zebrowitz, L. A. (2007). Do impressions of health, dominance, and warmth explain why masculine faces are preferred more in a short-term mate? *Evolutionary Psychology*, *5*(1), 15–27. doi: 10.1177/147470490700500102

Lui, Y. M., Bolme, D., Draper, B. A., Beveridge, J. R., Givens, G., & Phillips, P. J. (2009). A meta-analysis of face recognition covariates. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on* (pp. 1–8). IEEE.

Macrae, C. N., & Martin, D. (2007). A boy primed Sue: Feature-based processing and person construal. *European Journal of Social Psychology, 37*(5), 793–805. doi: 10.1002/ejsp.406

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science, 28*(2), 209–226. https://doi.org/10.1016/j.cogsci.2003.11.004

Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, *3*(2), 215–221. doi:10.3758/BF03212421

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'hello'? Personality impressions from brief novel voices. *PLoS One, 9*, e90779. https://doi.org/10.1371/journal.pone.0090779

McCaffery, J. M., & Burton, A. M. (2016). Passport checks: Interactions between matching faces and biographical details. *Applied Cognitive Psychology*, *30*(6), 925–933. doi: 10.1002/acp.3281

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*(1), 81–90. http://dx.doi.org/10.1037/0022-3514.52.1.81

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748. doi: 10.1038/264746a0

McIntyre, A. H., Hancock, P. J., Kittler, J., & Langton, S. R. (2013). Improving discrimination and face matching with caricature. *Applied Cognitive Psychology, 27*(6), 725–734. doi: 10.1002/acp.2966

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865–876. doi: 10.3758/BF03193433

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics*, *69*(7), 1175–1184. doi: 10.3758/BF03193954

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance on eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14*(4), 364–372. http://dx.doi.org/10.1037/a0013464

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching Face Images Taken on the Same Day or Months Apart: The Limitations of Photo ID. *Applied Cognitive Psychology, 27*(6), 700–706. doi: 10.1002/acp.2965

Mehrabian, A., & Ferris, S. R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology, 31*(3), 248–252. http://dx.doi.org/10.1037/h0024648

Melmoth, D. R., Kukkonen, H. T., Ma̋kela̋, P. K., & Rovamo, J. M. (2000). The effect of contrast and size scaling on face perception in foveal and extrafoveal vision. *Investigative Ophthalmology & Visual Science, 41*(9), 2811–2819.

Menzel, C., Hayn-Leichsenring, G. U., Langner, O., Wiese, H., & Redies, C. (2015). Fourier power spectrum characteristics of face photographs: Attractiveness perception depends on low-level image properties. *PloS One, 10*(4), e0122801. https://doi.org/10.1371/journal.pone.0122801

Michel, C., Corneille, O., & Rossion, B. (2007). Race categorization modulates holistic face encoding. *Cognitive Science, 31*(5), 911–924. doi: 10.1080/03640210701530805

Mileva, V. R., Jones, A. L., Russell, R., & Little, A. C. (2016). Sex differences in the perceived dominance and prestige of women with and without cosmetics. *Perception, 45*(10), 1166–1183. doi: 10.1177/0301006616652053

Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior, 27*(4), 237–254. doi: 10.1023/A:1027332800296

Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: The salience and significance of age in social judgments. *Advances in*

Experimental Social Psychology, *30*, 93–161. https://doi.org/10.1016/S0065-2601(08)60383-4

Montepare, J. M., & Zebrowitz, L. A. (2002). A social-developmental view of ageism. In T. D. Nelson (Ed.), *Ageism: Stereotyping and prejudice against older persons* (pp. 77–125). Cambridge, MA: The MIT Press.

Montepare, J. M., & Zebrowitz-McArthur, L. (1987). Perceptions of adults with childlike voices in two cultures. *Journal of Experimental Social Psychology*, *23*(4), 331–349. https://doi.org/10.1016/0022-1031(87)90045-X

Montepare, J. M., & Zebrowitz-McArthur, L. (1989). Children's perceptions of babyfaced adults. *Perceptual and Motor Skills*, *69*(2), 467–472. 10.2466/pms.1989.69.2.467

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist*, *111*(981), 855–869. doi: 10.1086/283219

Mueller, J. H., Heesacker, M., & Ross, M. J. (1984). Likability of targets and distractors in facial recognition. *American Journal of Psychology, 97*(2), 235–247. doi: 10.2307/1422598

Mueller, J. H., Thompson, W. B., & Vogel, J. M. (1988). Perceived honesty and face memory. *Personality and Social Psychology Bulletin, 14*(1), 114–124. doi: 10.1177/0146167288141012

Mulford, M., Orbell, J., Shatto, C., & Stockard, J. (1998). Physical attractiveness, opportunity, and success in everyday exchange. *American Journal of Sociology*, *103*(6), 1565–1592. doi: 10.1086/231401

Muscarella, F., & Cunningham, M. R. (1996). The evolutionary significance and social perception of male pattern baldness and facial hair. *Ethology and Sociobiology*, *17*(2), 99–117. https://doi.org/10.1016/0162-3095(95)00130-1

Na, J., Kim, S., Oh, H., Choi, I., & O'Toole, A. (2015). Competence judgments based on facial appearance are better predictors of American elections

than of Korean elections. *Psychological Science, 26*(7), 1107–1113. doi: 10.1177/0956797615576489

Nelson, D. L., Reed, V. S., & McEvoy, C. L. (1977). Learning to order pictures and words: A model of sensory and semantic encoding. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(5), 485–497. http://dx.doi.org/10.1037/0278-7393.3.5.485

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition, 128*(1), 56–63. https://doi.org/10.1016/j.cognition.2013.03.006

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., & Abdi, H. (1994). Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, *22*(2), 208–224. doi: 10.3758/BF03208892

O'Toole, A. J., Edelman, S., & Bulthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research, 38*(15), 2351–2363. http://dx.doi.org/10.1016/S0042-6989(98)00042-X

Ohala, J. J. (1982). The voice of dominance. *The Journal of the Acoustical Society of America, 72*(S1), S66. http://dx.doi.org/10.1121/1.2020007

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, *41*(1), 1–16. doi: 10.1159/000261706

Ohala, J. J., Hinton, L., & Nichols, J. (1997). Sound symbolism. In *Proc. 4th Seoul International Conference on Linguistics [SICOL]* (pp. 98–103).

Oldmeadow, J. A., Sutherland, C. A. M., & Young, A. W. (2013). Facial stereotype visualization through image averaging. *Social Psychological and Personality Science, 4*(5), 615–623. doi: 10.1177/1948550612469820

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570. https://doi.org/10.1016/j.tics.2014.09.007

Olivola, C. Y., Sussman, A. B., Tsetsos, K., Kang, O. E., & Todorov, A. (2012). Republicans prefer Republican-looking leaders: Political facial stereotypes predict candidate electoral success among right-leaning voters. *Social Psychological and Personality Science, 3*(5), 605–613. doi: 10.1177/1948550611432770

Olivola, C. Y., & Todorov, A. (2010a). Elected in 100 milliseconds: appearance-based trait inferences and voting. *Journal of Nonverbal Behaviour, 34*(2), 83–110. doi: 10.1007/s10919-009-0082-1

Olivola, C. Y., & Todorov, A. (2010b). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, *46*(2), 315–324. https://doi.org/10.1016/j.jesp.2009.12.002

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

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

Özener, B. (2011). Facial width-to-height ratio in a Turkish population is not sexually dimorphic and is unrelated to aggressive behavior. *Evolution and Human Behavior*, 33(3), 169–173. doi: 10.1016/j.evolhumbehav.2011.08.001.

Parry, F. M., Young, A. W., Shona, J., Saul, M., & Moss, A. (1991). Dissociable face processing impairments after brain injury. *Journal of Clinical and Experimental Neuropsychology*, *13*(4), 545–558. http://dx.doi.org/10.1080/01688639108401070

Pazda, A. D., Thorstenson, C. A., Elliot, A. J., & Perrett, D. I. (2016). Women's facial redness increases their perceived attractiveness: Mediation through perceived healthiness. *Perception*, *45*(7), 739–754. doi: 10.1177/0301006616633386

Pear, T. H. (1931). *Voice and personality*. London: Chapman and Hall.

Peli, E. (1990). Contrast in complex images. *JOSA A*, *7*(10), 2032–2040. https://doi.org/10.1364/JOSAA.7.002032

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. doi: 10.1163/156856897X00366

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*(2), 169–178. doi: 10.1007/s004420100720

Perrett, D. I. (1994). Facial shape and judgements. *Nature*, *368*(6468), 239–242. http://dx.doi.org/10.1038/368239a0

Perrett, D. I. (2010). *In your face: The new science of human attraction*. Basingstoke, UK: Palgrave Macmillan.

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., ... & Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature*, *394*(6696), 884–887. doi: 10.1038/29772

Peskin, M., & Newell, F. N. (2004). Familiarity breeds attraction: Effect of exposure on the attractiveness of typical and distinctive faces. *Perception, 33*(2), 147–157. doi: 10.1068/p5028

Peters, M., Rhodes, G., & Simmons, L. W. (2008). Does attractiveness in men provide clues to semen quality? *Journal of Evolutionary Biology*, *21*(2), 572–579. doi:10.1111/j.1420-9101.2007.01477.x

Plataniotis, K., & Venetsanopoulos, A. N. (2013). *Color image processing and applications*. Springer Science & Business Media.

Porcheron, A., Mauger, E., & Russell, R. (2013). Aspects of facial contrast decrease with age and are cues for age perception. *PLoS One, 8*(3), e57985. http://doi.org/10.1371/journal.pone.0057985

Porter, S., England, L., Juodis, M., ten Brinke, L., & Wilson, K. (2008). Is the face a window to the soul? Investigation of the accuracy of intuitive judgments of the trustworthiness of human faces. *Canadian Journal of Behavioural Science, 40*(3), 171–177. doi: 10.1037/0008-400X.40.3.171

Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly, 26*(4), 269–281. doi: 10.1111/1471-6402.t01-1-00066

Puhl, R. M., & Heuer, C. A. (2009). The stigma of obesity: A review and update. *Obesity*, *17*(5), 941–964. doi: 10.1038/oby.2008.636

Puts, D. A. (2005). Mating context and menstrual phase affect women's preferences for male voice pitch. *Evolution and Human Behavior*, *26*(5), 388–397. http://dx.doi.org/10.1016/j.evolhumbehav.2005.03.001

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior, 27*(4), 283–296. http://dx.doi.org/10.1016/j.evolhumbehav.2005.11.003

Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. (2007). Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, *28*(5), 340–344. http://dx.doi.org/10.1016/j.evolhumbehav.2007.05.002

Rainis, N. (2001). Semantic contexts and face recognition. *Applied Cognitive Psychology*, *15*(2), 173–186. doi: 10.1002/1099-0720(200103/04)15:2<173::AID-ACP695>3.0.CO;2-Q

Rakić, T., Steffens, M. C., & Mummendey, A. (2011). Blinded by the accent! The minor role of looks in ethnic categorization. *Journal of Personality and Social Psychology*, *100*(1), 16–29. http://dx.doi.org/10.1037/a0021522

Ramsey, J. L., Langlois, J. H., Hoss, R. A., Rubenstein, A. J., & Griffin, A. M. (2004). Origins of a stereotype: Categorization of facial attractiveness by 6-month-old infants. *Developmental Science*, *7*(2), 201–211. doi: 10.1111/j.1467-7687.2004.00339.x

Re, D. E., Whitehead, R. D., Xiao, D., & Perrett, D. I. (2011). Oxygenated-blood colour change thresholds for perceived facial redness, health,

and attractiveness. *PloS One, 6*(3), e17859. https://doi.org/10.1371/journal.pone.0017859

Redies, C., Hänisch, J., Blickhan, M., & Denzler, J. (2007). Artists portray human faces with the Fourier statistics of complex natural scenes. *Network: Computation in Neural Systems*, *18*(3), 235–248. http://dx.doi.org/10.1080/09548980701574496

Redies, C., Hasenstein, J., & Denzler, J. (2007). Fractal-like image statistics in visual art: Similarity to natural scenes. *Spatial Vision, 21*(1), 137–148. doi: 10.1163/156856807782753921

Rezlescu, C., Penton, T., Walsh, V., Tsujimura, H., Scott, S. K., & Banissy, M. J. (2015). Dominant voices and attractive faces: The contribution of visual and auditory information to integrated person impressions. *Journal of Nonverbal Behavior*, *39*(4), 355–370. doi: 10.1007/s10919-015-0214-8

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

Rhodes, G., Halberstadt, J., & Brajkovich, G. (2001). Generalization of mere exposure effects to averaged composite faces. *Social Cognition, 19*(1), 57–70. doi: 10.1521/soco.19.1.57.18961

Rhodes, G., Halberstadt, J., Jeffery, L., & Palermo, R. (2005). The attractiveness of average faces is not a generalized mere exposure effect. *Social Cognition*, *23*(3), 205–217. doi: 10.1521/soco.2005.23.3.205

Rice, A., Phillips, P. J., Natu, V., An, X., & O'Toole, A. J. (2013). Unaware person recognition from the body when face identification fails. *Psychological Science*, *24*(11), 2235–2243. doi: 10.1177/0956797613492986

Rice, A., Phillips, P. J., & O'Toole, A. (2013). The role of the face and body in unfamiliar person identification. *Applied Cognitive Psychology*, *27*(6), 761–768. doi:10.1002/acp.2969

Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science, 28*(2), 147–166. https://doi.org/10.1016/j.cogsci.2003.11.003

Robinson, K., Blais, C., Duncan, J., Forget, H., & Fiset, D. (2014). The dual nature of the human face: there is a little Jekyll and a little Hyde in all of us. *Frontiers in Psychology*, *5*, 139. doi: doi: 10.3389/fpsyg.2014.00139

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters*, *73*(6), 814–817. doi: 10.1103/PhysRevLett.73.814

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57*(4), 743–762. doi: 10.1111/0022-4537.00239

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

Rule, N. O., & Ambady, N. (2010). Democrats and Republicans can be differentiated from their faces. *PLoS One, 5*(1), e8733. http://dx.doi.org/10.1371/journal.pone.0008733

Rule, N. O., Ambady, N., & Hallett, K. C. (2009). Female sexual orientation is perceived accurately, rapidly, and automatically from the face and its features. *Journal Experimental Social Psychology, 45*(6), 1245–1251. http://dx.doi.org/10.1016/j.jesp.2009.07.010

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioural and neural correlates. *Journal of Personality and Social Psychology, 104*(3), 409–426. http://dx.doi.org/10.1037/a0031050

Russell, R. (2003). Sex, beauty, and the relative luminance of facial features. *Perception, 32*(9), 1093–1107. http://doi.org/10.1068/p5101

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

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

Saxton, T. K., Burriss, R. P., Murray, A. K., Rowland, H. M., & Roberts, S. C. (2009). Face, body and speech cues independently predict judgments of attractiveness. *Journal of Evolutionary Psychology*, *7*(1), 23–35. http://dx.doi.org/10.1556/JEP.7.2009.1.4

Scheib, J. E., Gangestad, S. W., & Thornhill, R. (1999). Facial attractiveness, symmetry and cues of good genes. *Proceedings of the Royal Society of London B: Biological Sciences*, *266*(1431), 1913–1917. doi: 10.1098/rspb.1999.0866

Schneider, D. J. (1973). Implicit personality theory: Review. *Psychological Bulletin, 79*(5), 294–309. http://dx.doi.org/10.1037/h0034496

Schweinberger, S. R., Herholz, A., & Stief, V. (1997). Auditory long term memory: Repetition priming of voice recognition. *The Quarterly Journal of Experimental Psychology: Section A*, *50*(3), 498–517. http://dx.doi.org/10.1080/713755724

Schweinberger, S. R., Kloth, N., & Robertson, D. M. (2011). Hearing facial identities: Brain correlates of face–voice integration in person identification. *Cortex*, *47*(9), 1026–1037. http://dx.doi.org/10.1080/17470210601063589

Schweinberger, S. R., Robertson, D., & Kaufmann, J. M. (2007). Hearing facial identities. *The Quarterly Journal of Experimental Psychology*, *60*(10), 1446–1456. http://dx.doi.org/10.1080/17470210601063589

Sczesny, S., Spreemann, S., & Stahlberg, D. (2006). Masculine = competent? Physical appearance and sex as sources of gender-stereotypic attributions. *Swiss Journal of Psychology, 65*(1), 15–23. doi: 10.1024/1421-0185.65.1.15

Searle, J. R. (1979). *Expression and meaning: Studies in the theory of speech acts*. Cambridge: Cambridge University Press.

Secord, P. F. (1958). Facial features and inference processes in interpersonal perception. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior* (pp. 300–315). Stanford, CA: Stanford University Press.

Secord, P. F. (1959). Stereotyping and favorableness in the perception of Negro faces. *The Journal of Abnormal and Social Psychology*, *59*(3), 309–314. http://dx.doi.org/10.1037/h0042001

Secord, P. F., & Jourard, S. M. (1956). Mother-concepts and judgments of young women's faces. *The Journal of Abnormal and Social Psychology*, *52*(2), 246–250. http://dx.doi.org/10.1037/h0048054

Shan, S., Gao, W., & Zhao, D. (2003). Face recognition based on face-specific subspace. *International Journal of Imaging Systems and Technology*, *13*(1), 23–32. doi: 10.1002/ima.10047

Shimamura, A. P., Ross, J. G., & Bennett, H. D. (2006). Memory for facial expressions: The power of a smile. *Psychonomic Bulletin & Review*, *13*(2), 217–222. doi: 10.3758/BF03193833

Shriver, E. R., Young, S. G., Hugenberg, K., Bernstein, M. J., & Lanter, J. R. (2008). Class, race, and the face: Social context modulates the cross-race effect in face recognition. *Personality and Social Psychology Bulletin*, *34*(2), 260–274. doi: 10.1177/0146167207310455

Siibak, A. (2009). Constructing the self through the photo selection-visual impression management on social networking websites. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 3*(1), article 1.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. http://dx.doi.org/10.1037/0096-3445.117.1.34

Sofer, C., Dotsch, R., Wigboldus, D. H., & Todorov, A. (2015). What is typical is good: The influence of face typicality on perceived

trustworthiness. *Psychological Science*, *26*(1), 39–47. doi: 10.1177/0956797614554955

Solomon, J. A. (2002). Noise reveals visual mechanisms of detection and discrimination. *Journal of Vision, 2*(1), 105–120. doi: 10.1167/2.1.7

Song, H., Vonasch, A. J., Meier, B. P., & Bargh, J. A. (2012). Brighten up: Smiles facilitate perceptual judgment of facial lightness. *Journal of Experimental Social Psychology*, *48*(1), 450–452. https://doi.org/10.1016/j.jesp.2011.10.003

Stephen, I. D., Coetzee, V., & Perrett, D. I. (2011). Carotenoid and melanin pigment coloration affect perceived human health. *Evolution and Human Behavior, 32*(3), 216–227. https://doi.org/10.1016/j.evolhumbehav.2010.09.003

Stephen, I. D., Coetzee, V., Smith, M. L., & Perrett, D. I. (2009). Skin blood perfusion and oxygenation colour affect perceived human health. *PloS One, 4*(4), e5083. https://doi.org/10.1371/journal.pone.0005083

Stephen, I. D., Oldham, F. H., Perrett, D. I., & Barton, R. A. (2012). Redness enhances perceived aggression, dominance and attractiveness in men's faces. *Evolutionary Psychology*, *10*(3), 147470491201000312. doi: 10.1177/147470491201000312

Stephen, I. D., & Perrett, D. I. (2015). Color and face perception. In A. Elliot, M. Fairchild, & A. Franklin (Eds.), *Handbook of color psychology* (pp. 585–602). Cambridge, UK: Cambridge University Press.

Stephen, I. D., Smith, M. L., Stirrat, M. R., & Perrett, D. I. (2009). Facial skin coloration affects perceived health of human faces. *International Journal of Primatology, 30*(6), 845–857. doi: 10.1007/s10764-009-9380-z

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

Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica, 36*, 314–331.

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*(1), 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. doi: 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*(2), 186–208. doi: 10.1111/bjop.12085

Sutherland, C. A. M., Young, A. W., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology, 108*(2), 397–415. doi: 10.1111/bjop.12206

Tan, K. W., & Stephen, I. D. (2013). Colour detection thresholds in faces and colour patches. *Perception, 42*(7), 733–741. doi: 10.1068/p7499

Thornhill, R., & Gangestad, S. W. (1993). Human facial beauty: Averageness, symmetry, and parasite resistance. *Human Nature*, *4*(3), 237–269. doi: 10.1007/BF02692201

Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in Cognitive Science*, *3*(12), 452–460. https://doi.org/10.1016/S1364-6613(99)01403-5

Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., & Feinberg, D. R. (2012) Voice pitch influences voting behavior. *Evolutionary Behavioral Sciences, 5*(4), 323

Titze, I. R., & Martin, D. W. (1998). Principles of voice production. *The Journal of the Acoustical Society of America*, *104*(3), 1148–1148. doi: http://dx.doi.org/10.1121/1.424266

Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signalling approach/avoidance behaviours. *Annals of the New York Academy of Sciences, 1124*(1), 208–224. doi: 10.1196/annals.1440.012

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*(4), 724–738. http://dx.doi.org/10.1037/a0032335

Todorov, A., Dotsch, R., Wigboldus, D. H. J., & Said, C. P. (2011). Data-driven methods for modeling social perception. *Social and Personality Psychology Compass, 5*(10), 775–791. doi: 10.1111/ j.1751-9004.2011.00389.x

Todorov, A., Loehr, V., & Oosterhof, N. N. (2010). The obligatory nature of holistic processing of faces in social judgments. *Perception, 39*(4), 514–532. doi: 10.1068/p6501

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623–1626. doi: 10.1126/science.1110589

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

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

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*(6), 813–833. doi: 10.1521/soco.2009.27.6.813

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different images of the same person. *Psychological Science, 25*(7), 1404–1417. doi: 10.1177/0956797614532474

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

Tomlin, R. J., Stevenage, S. V., & Hammond, S. (2016). Putting the pieces together: Revealing face–voice integration through the facial overshadowing effect. *Visual Cognition*. Advance online publication. http://d.doi.org/10.1080/13506285.2016.1245230

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception*, *43*(2), 214–218. doi: 10.1068/p7676

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, *23*(1), 47–58. http://dx.doi.org/10.1037/xap0000108

Traunmüller, H., & Eriksson, A. (1995). *The frequency range of the voice fundamental in the speech of male and female adults*. Unpublished manuscript. [Available at http://www2.ling.su.se/staff/hartmut/f0_m&f.pdf].

Tsankova, E., Krumhuber, E., Aubrey, A. J., Kappas, A., Möllering, G., Marshall, D., & Rosin, P. L. (2015). The multi-modal nature of trustworthiness perception. In *AVSP* (pp. 147–152).

Tsantani, M. S., Belin, P., Paterson, H. M., & McAleer, P. (2016). Low vocal pitch preference drives first impressions irrespective of context in male voices but not in female voices. *Perception*, *45*(8), 946–463. doi: 10.1177/0301006616643675

Turk, M. A., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*(1), 71–86. doi: 10.1109/CVPR.1991.139758

Tusing, K. J., & Dillard, J. P. (2000). The sounds of dominance. *Human Communication Research*, *26*(1), 148–171. doi: 10.1111/j.1468-2958.2000.tb00754.x

Valdez, P., & Mehrabian, A. (1994). Effects of color on emotions. *Journal of Experimental Psychology. General*, *123*(4), 394–409. doi: 10.1037/0096-3445.123.4.394

Verosky, S. C., & Todorov, A. (2010a). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science*, *21*(6), 779–785. doi: 10.1177/0956797610371965

Verosky, S. C., & Todorov, A. (2010b). Differential neural responses to faces physically similar to the self as a function of their valence. *NeuroImage*, *49*(2), 1690–1698. https://doi.org/10.1016/j.neuroimage.2009.10.017

Verosky, S. C., & Todorov, A. (2013). When physical similarity matters: Mechanisms underlying affective learning generalization to the evaluation of novel faces. *Journal of Experimental Social Psychology*, *49*(4), 661–669. https://doi.org/10.1016/j.jesp.2013.02.004

Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: The sequel: A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, *28*(4), 260–271. https://doi.org/10.1016/j.evolhumbehav.2007.04.006

Vukovic, J., Feinberg, D. R., Jones, B. C., DeBruine, L. M., Welling, L. L. M., Little, A. C., & Smith, F. G. (2008). Self-rated attractiveness predicts individual differences in women's preferences for masculine men's voices. *Personality and Individual Differences*, *45*(6), 451–456. http://dx.doi.org/10.1016/j.paid.2008.05.013

Vukovic, J., Jones, B. C., Feinberg, D. R., DeBruine, L. M., Smith, F. G., Welling, L. L., & Little, A. C. (2011). Variation in perceptions of physical dominance and trustworthiness predicts individual differences in the effect of relationship context on women's preferences for masculine

pitch in men's voices. *British Journal of Psychology, 102*(1), 37–48. doi: 10.1348/000712610X498750

Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision, 9*(11), 1–13. doi: 10.1167/9.11.12.

Walker, M., & Vetter, T. (2015). Changing the personality of a face: Perceived big two and big five personality factors modelled in real photographs. *Journal of Personality and Social Psychology*, *110*(4), 609–624. http://dx.doi.org/10.1037/pspp0000064

Warren, C., & Morton, J. (1982). The effects of priming on picture recognition. *British Journal of Psychology*, *73*(1), 117–129. doi: 10.1111/j.2044-8295.1982.tb01796.x

Weeden, J., & Sabini, J. (2005). Physical attractiveness and health in Western societies: A review. *Psychological Bulletin, 131*(5), 645–653. http://dx.doi.org/10.1037/0033-2909.131.5.635

Wells, T. J., Dunn, A. K., Sergeant, M. J. T., & Davies, M. N. O. (2009). Multiple signals in human mate selection: A review and framework for integrating facial and vocal signals. *Journal of Evolutionary Psychology*, *7*(2), 111–139. http://dx.doi.org/10.1556/JEP.7.2009.2.2

Wen, C. Y., & Chou, C. M. (2004). Color Image Models and its Applications to Document Examination. *Forensic Science Journal*, *3*(1), 23–32.

White, D., Burton, A. M., Jenkins, R., & Kemp, R. I. (2014a). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied, 20*(2), 166–173. doi: 10.1037/xap0000009

White, D., Burton, A. L., & Kemp, R. I. (2015). Not looking yourself: The cost of self-selecting photographs for identity verification. *British Journal of Psychology, 107*(2), 359–373. doi: 10.1111/bjop.12141

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied cognitive psychology*, *27*(6), 769–777. doi: 10.1002/acp.2971

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014b). Feedback training for facial image comparison. *Psychonomic Bulletin and Review, 21*(1), 100–106. doi: 10.3758/s13423-013-0475-3

White, D., Sutherland, C. A. M., & Burton, A. L. (2017). Choosing face: The curse of self in profile image selection. *Cognitive Research: Principles and Implications, 2*, 23. doi: 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*(3), 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*(7), 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*(4), 331–338. doi: 10.1177/1948550615624142

Winograd, E. (1976). Recognition memory for faces following nine different judgments. *Bulletin of the Psychonomic Society*, *8*(6), 419–421. doi: 10.3758/BF03335185

Wolff, S. E., & Puts, D. A. (2010). Vocal masculinity is a robust dominance signal in men. *Behavioral Ecology and Sociobiology*, *64*(10), 1673–1683. doi: 10.1007/s00265-010-0981-5

Woodhead, M. M., Baddeley, A. D., & Simmonds, D. C. V. (1979). On Training People to Recognize Faces. *Ergonomics*, *22*(3), 333–343. http://dx.doi.org/10.1080/00140137908924617

Xu, Y., & Kelly, A. (2010). Perception of anger and happiness from resynthesized speech with size-related manipulations. In *Proceedings of the 5th International Conference on Speech Prosody (SP2010). Chicago, IL.*

Yacoob, Y., & Davis, L. (2002). Smiling faces are better for face recognition. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on* (pp. 59–64). IEEE.

Young, A. W., & Bruce, V. (2011). Understanding person perception. *British Journal of Psychology*, *102*(4), 959–974. doi: 10.1111/j.2044-8295.2011.02045.x

Young A. W., & Burton, A. M. (in press). Recognizing faces. *Current Directions in Psychological Science.*

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception*, *14*(6), 737–746. doi: 10.1068/p140737

Young, A. W., Hellawell, D., & Day, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747–759. doi: 10.1068/p160747n

Young, A. W., McWeeny, K. H., Hay, D. C., & Ellis, A. W. (1986). Matching familiar and unfamiliar faces on identity and expression. *Psychological Research*, *48*(2), 63–68. doi: 10.1007/BF00309318

Young, A. W., Newcombe, F., de Haan, E. H., & Hay, D. C. (1993). Face perception after brain injury. *Brain*, *116*(4), 941–959. https://doi.org/10.1093/brain/116.4.941

Young, S. G., Elliot, A. J., Feltman, R., & Ambady, N. (2013). Red enhances the processing of facial expressions of anger. *Emotion*, *13*(3), 380–384. http://dx.doi.org/10.1037/a0032471

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6), 263–271. https://doi.org/10.1016/j.tics.2013.04.004

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*(2), 1–27. http://dx.doi.org/10.1037/h0025848

Zebrowitz, L. A. (1996). Physical appearance as a basis for stereotyping. In M. H. N. McRae & C. Stangor (Eds.), *Foundation of stereotypes and stereotyping* (pp. 79–120). New York, NY: Guilford Press.

Zebrowitz, L. A. (1997). *Reading faces: Window to the soul?* Westview Press.

Zebrowitz, L. A. (2011). Ecological and social approaches to face perception. *The Oxford handbook of face perception*, 31–50.

Zebrowitz, L. A., Bronstad, P. M., & Lee, H. K. (2007). The contribution of face familiarity to ingroup favoritism and stereotyping. *Social Cognition*, *25*(2), 306–338. doi: 10.1521/soco.2007.25.2.306

Zebrowitz, L. A., & Collins, M. A. (1997). Accurate social perception at zero acquaintance: The affordances of a Gibsonian approach. *Personality and Social Psychology Review*, *1*(3), 204–223. doi: 10.1207/s15327957pspr0103_2

Zebrowitz, L. A., Fellous, J. M., Mignault, A., & Andreoletti, C. (2003). Trait impressions as overgeneralized responses to adaptively significant facial qualities: Evidence from connectionist modeling. *Personality and Social Psychology Review*, *7*(3), 194–215. doi: 10.1207/S15327957PSPR0703_01

Zebrowitz, L. A., Hall, J. A., Murphy, N. A., & Rhodes, G. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin*, *28*(2), 238–249. doi: 10.1177/0146167202282009

Zebrowitz, L. A., Kikuchi, M., & Fellous, J. M. (2010). Facial resemblance to emotions: Group differences, impression effects, and race stereotypes. *Journal of Personality and Social Psychology*, *98*(2), 175–189. http://dx.doi.org/10.1037/a0017990

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass, 2*(3), 1497–1517. doi: 10.1111/j.1751-9004.2008.00109.x

Zebrowitz, L. A., Montepare, J. M., & Lee, H. K. (1993). They don't all look alike: Individual impressions of other racial groups. *Journal of Personality and Social Psychology*, *65*(1), 85–101. http://dx.doi.org/10.1037/0022-3514.65.1.85

Zebrowitz, L. A., & Rhodes, G. (2004). Sensitivity to 'bad genes' and the anomalous face overgeneralization effect: Cue validity, cue utilization, and accuracy in judging intelligence and health. *Journal of Nonverbal Behavior*, *28*(3), 167–185. doi: 10.1023/B:JONB.0000039648.30935.1b

Zebrowitz, L. A., Wadlinger, H. A., Luevano, V. X., White, B. M., Xing, C., & Zhang, Y. (2011). Animal analogies in first impressions of faces. *Social Cognition*, *29*(4), 486–496. doi: 10.1521/soco.2011.29.4.486

Zebrowitz-McArthur, L., & Berry, D. S. (1987). Cross-cultural agreement in perceptions of babyfaced adults. *Journal of Cross Cultural Psychology, 18*(2), 165–192. doi: 10.1177/0022002187018002003

Zell, E., & Balcetis, E. (2012). The influence of social comparison on visual representation of one's face. *PLoS One, 7*(5), e36742. doi: 10.1371/journal.pone.0036742

Zuckerman, M., & Driver, R. E. (1989). What sounds beautiful is good - the vocal attractiveness stereotype. *Journal of Nonverbal Behavior, 13*(2), 67–82. doi: 10.1007/BF00990791

Zuckerman, M., Hodgins, H., & Miyake, K. (1990). The vocal attractiveness stereotype: Replication and elaboration. *Journal of Nonverbal Behavior, 14*(2), 97–112. doi: 10.1007/BF01670437

Zuckerman, M., & Miyake, K. (1993). The attractive voice: What makes it so? *Journal of Nonverbal Behavior*, *17*(2), 119–135. doi: 10.1007/BF01001960

Zuckerman, M., Miyake, K., & Elkin, C. S. (1995). Effects of attractiveness and maturity of face and voice on interpersonal impressions. *Journal of Research in Personality, 29*(2), 253–272. http://dx.doi.org/10.1006/jrpe.1995.1015

Zuckerman, M., Miyake, K., & Hodgins, H. S. (1991). Cross-channel effects of vocal and physical attractiveness and their implications for interpersonal perception. *Journal of Personality and Social Psychology, 60*(4), 545–554. http://dx.doi.org/10.1037/0022-3514.60.4.545