First impressions from faces: Integrating facial dimensions and social categories

Clare A.M. Sutherland

Doctor of Philosophy

University of York

Psychology

March 2015

To my family.

Abstract

When we meet a person for the first time, we can gain a wealth of information from perceiving their face - for example, their age, sex, ethnicity and level of attractiveness. However, we also make more subjective facial judgments of the character or personality of the person depicted as well; for example, by judging them on their trustworthiness or competence. Since these facial first impressions are linked to important real-life consequences, it is important that we have a solid theoretical understanding of which judgments are important and how they are made. Here, I first model the dimensions underlying first impressions made to highly variable, naturalistic photographs of faces and find three key dimensions: approachability, dominance and youthful attractiveness. I then examine how categorical facial information (such as the gender or culture of the face or perceiver) interacts with these key facial impression dimensions, finding that there are differences in the structure and evaluation of trait impressions by face and perceiver gender and culture. My findings demonstrate that while dimensional approaches have brought a useful guiding framework to the field of facial first impressions, categorical social groups also need to be included within these models. Specifically, while the functionality of key facial impression dimensions may be similar across social groups, the cues or specific traits underlying these dimensions seem to differ depending on the social group of the perceiver or target face.

Table of contents

List of tables

List of figures

List of supplementary material

Acknowledgements

Thank you Andy and Julian, for such great supervision. I expected to be challenged and to think hard in our meetings but I didn't anticipate just how much I would enjoy them as well. Also thank you both for your constant enthusiasm and for your unfailing support. I could not have asked for better supervisors.

I am grateful to the ESRC for all of their financial support; also to the EPS, Brain, the Rank Prize funds and PsyPAG for funding dissemination of my research.

I would never have got this far without the friendly and supportive environment at York. Mladen, Tessa, Dave, Rich and Xiaoqian, thank you for the countless times you helped me; Sally, I really enjoyed our conversations on social psychology. Also to Rob and Tom, my TAP committee and the members of Facelab and Social Minds for all of the useful suggestions and great feedback along the way. There were also many students who deserve thanks for their excellent research assistance: especially, I would like to thank Carmen for the coding help, and to Celia, Kristy and Lindsay for all of your cultural knowledge and your critical thoughts, which helped shape my thinking on the cross-cultural work. I had fun learning about Chinese views on first impressions! There were so many people at York who helped guide me; thank you all.

To my friends at York and elsewhere – a big thank you for all your support. Especially to the bake off group (Adele, Mark, Dave, Alex, Sarah, Emma, Eilidh and Charlotte) for the games, crazy adventures and for all of the great meals! To Elaine and Xin for helping me explore Yorkshire; I really enjoyed our trips. Fraser, Emma, Lyndsay, Eni, Michael, and Gaius thank you for listening when I needed it, and for looking over drafts of papers (sorry about that….) Jiajun, I enjoyed our times in cafes or the library (and the dim sum of course!). Thank you all for making these years so special and so much fun.

Of course, I would like to thank my family for always being there when I needed to hear a voice from home. You shared the highs and lows, encouraged me to dream big and to follow my heart, carefully proofread my work, and gave me much-needed pep talks. "Thumps up!". Finally, especially to Scott, for all your love and support, for your careful proofreading and formatting advice, and for reminding me why I love science.

Author's declaration

This thesis was written by myself and represents my original work, with supervision from Professor Andrew Young. The research was funded by a quota 1+3 studentship from the ESRC.

Chapter two: The ratings data in Experiment 2, were collected by myself as part of my MRes in Psychology at the University of York (2011), and also by Isobel Santos and John Towler for their PhD and MRes respectively. I re-analysed these data here. Chapter two also appears here as it was accepted for publication in Cognition (Sutherland et al., 2013). It represents my original work: I wrote the first draft; the other authors, the editor and two three reviewers suggested modifications, and I chose which revisions were made. Portions of these data were also presented at the White Rose doctoral training conference, York, 2012.

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118. doi:10.1016/j.cognition.2012.12.001

Chapter three: appears here as it will be submitted to a journal (Sutherland, Oldmeadow, & Young, in prep). It represents my original work; I wrote the first draft; the other authors suggested modifications and I chose which revisions were made. Portions of these data were also presented at the Visual Sciences Society meeting, Florida, 2014 and the European Conference of Visual Perception, Belgrade, 2014.

Sutherland, C. A. M., Oldmeadow, J. A., & Young, A. W. (in prep). Social and facial models of person perception show more similarity on the first than second dimension.

Chapter four: John Towler and Philippa Howsley collected pilot data as part of their Masters in Research Psychology which was used to choose the stimuli used in Experiment 2; Carmen Mootz second coded the data in Studies 1 and 2. Chapter four also appears here as it was accepted for publication in the British Journal of Psychology (Sutherland, Young, Mootz, & Oldmeadow, in press). It represents my original work: I wrote the first draft; the other authors, the editor and two reviewers suggested

modifications, and I chose which revisions were made. Portions of these data were also presented at the Rank Prize meeting on Face Perception, Grasmere, March 2014.

Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (in press). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology*.

Chapter five: Mark Hymers helped programme the first experiment in Python; Celia Liu, Kristy Chu and Lingshan Zhang helped collect and sort the face stimuli, translated the stimuli and recruited and tested the majority of participants in both experiments. They also second coded the Chinese participant results. Xin Kang, Xiaoqian Yan, and Rosa Kwok helped sort the face stimuli, and helped translate the stimuli; finally Hollie Bryne and Josh Spowage second coded the British participant results. Portions of these data were presented at the Mind Across Cultures symposium, York, 2014, and the Society for Personality and Social Psychology meeting, California, 2015.

Chapter 1. Introduction

1.1. Overview

When we encounter a stranger for the first time, in real life or online, one of the most salient sources of information about that person is their face. From their face, perceivers can form an impression of the target's gender, age, ethnicity and attractiveness (Bruce & Young, 2012). Interestingly, people go further than these relatively objective judgements, and also spontaneously and readily judge a person's character and personality from their facial features (Hassin & Trope, 2000; Oosterhof & Todorov, 2008). Outside of the laboratory, the success of physiognomy throughout the 18th-20th centuries also attests to the historical popularity of reading a person's character from their face (cf. Hassin & Trope, 2000). This popularity has not diminished today, judging by the 5,560,000 Google hits for 'facial first impressions' (Google search, 05.01.15), which variously promise to give potential clients insights into their personality, optimal career and innermost secrets, all through the medium of their facial appearance. Moreover, a recent poll I conducted of 97 undergraduate students showed that around half (51%) agreed that you could tell something about someone's personality from their face alone (December 2013, Department of Psychology, University of York).

This review will define facial first impressions as *'rapidly made, far-reaching social inferences from visual facial cues'*, loosely borrowing from a definition of person perception by Quadflieg and Macrae (2011). Research has shown that these facial social inferences are fairly consistent across raters (Little & Perrett, 2007; Oosterhof & Todorov, 2008; Santos & Young, 2005, 2008, 2011; Todorov, Said, & Verosky, 2011) and that they seem to occur fairly rapidly (Willis & Todorov, 2006). For example, judgements of trustworthiness, attractiveness, likeability, competence and aggressiveness made to faces presented for as briefly as 100ms, considerably agreed with judgements made under free viewing conditions (Willis & Todorov, 2006). Judgements of trustworthiness made to faces viewed for only 33ms also showed considerable overlap with judgements made without time constraint (Todorov, Pakrashi, & Oosterhof, 2009). This suggests that we form our facial impressions remarkably

quickly and that these impressions are therefore fairly automatic (Todorov et al., 2009; Willis & Todorov, 2006).

This review will first briefly make the case for why it is important that we understand these facial first impressions, including what the consequences of these impressions are and how accurate they are. Second, face perception research into these facial judgements will be examined, starting with cues thought to subserve specific judgements and then considering a recent face perception model that synthesises these disparate findings. This section will also include a brief discussion of a recent, more ecological strategy of face sampling: the 'ambient image' approach. Third, social psychological research regarding social categorisation and stereotyping will be discussed, along with more recent attempts to integrate social stereotyping with face perception. Finally, this review will touch on links between social psychological studies and facial first impressions that have yet to be explored, and lay out the questions considered in the rest of the thesis.

While there is also an extensive literature on impressions made from real-life encounters or from short naturalistic videos (so-called 'thin slice' judgements: Ambady, Bernieri, & Richeson, 2000; Ambady et al., 2000; Ambady & Rosenthal, 1992; Naumann, Vazire, Rentfrow, & Gosling, 2009; Wall, Taylor, Dixon, Conchie, & Ellis, 2013), this thesis will concentrate only on impressions made from facial photographs. This is because facial impressions from photographs have been shown to have widespread real life consequences and online encounters often only encompass facial photographs, as will be discussed next. Moreover, there is also an extensive theoretical literature on facial impressions made from photographs, as will be discussed in the main body of the review.

1.2. The importance of facial first impressions

From the examples given in the introduction, it is clear that there is a strong public interest in facial first impressions that dates back to antiquity. Perhaps more profoundly, however, these judgements can also have important real-life consequences (see Olivola, Funk, & Todorov, 2014; and Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015 for

recent reviews). For instance, the perceived competence of facial photographs of politicians was found to predict election outcomes in the US at a rate higher than chance (Todorov, Mandisodza, Goren, & Hall, 2005). Indeed, nearly 70% of Senate races were correctly predicted in this way, ahead of the actual political outcome. Moreover, there was a correlation of .4 on average between the proportion of facial competence votes from participants in the study with the proportion of actual political votes for these candidates in the real election (Todorov et al., 2005). This finding extends across cultures (Antonakis & Dalgas, 2009; Lawson, Lenz, Baker, & Myers, 2010; Mattes & Milazzo, 2014) and strikingly, even leadership judgements of politicians' faces made by children predicted real election results (Antonakis & Dalgas, 2009).

While the influence of facial impressions on election results may be an especially surprising finding, there is no shortage of other examples of facial first impressions potentially influencing a wide range of important real-life events. For instance, inferences of leadership from face photographs track company profits and/or individual CEO compensation (Rule & Ambady, 2008, 2009; Rule, Ishii, & Ambady, 2011); impressions of competence and dominance predict job responsibilities (i.e. assets: Walker Naylor, 2007); and impressions of facial dominance made from photographs of cadets predicted their military rank more than twenty years later (Mueller & Mazur, 1996). Inferences of trustworthiness from face photographs also predict real online financial decisions (Duarte, Siegel, & Young, 2012; X. Yang, 2014) and litigants' facial babyfacedness can seemingly influence financial awards in real court decisions (Zebrowitz & McDonald, 1991). Finally, there is a long history of research showing that facial attractiveness affects a range of employment decisions, from initial hiring decisions to a career advancement (Gilmore, Beehr, & Love, 1986; Hochschild & Borch, 2011; Lutz, 2010; see Hosoda, Stone-Romero, & Coats, 2003 for a meta-analysis). While most of the previous findings are necessarily correlational, since they examine effects in the real world, casual links have also been demonstrated in experiments which manipulate faces on perceived social judgements, and then measure the effect of these facial judgements on important outcomes such as legal, political or financial decisions (Little, Burriss, Jones, & Verosky, 2007; Porter, ten Brinke, & Gustaw, 2010; Rezlescu, Duchaine, Olivola, & Chater, 2012; see Todorov et al., 2015 for a review).

The implications of facial first impressions from photographs will likely only become more widespread in future, since increasing numbers of people are socialising, trading, and dating with strangers online. To put this in context, by 2013, Facebook had over 1.23 billion active users (Sedghi, 2014) and LinkedIn had over 178 million active users (Quantcast, 2014); while half of British adults currently searching for a relationship have used online dating (YouGov, 2014). Since initial online encounters are often restricted to the visual information provided by one or a few main profile photographs, the consequences of facial first impressions from photographs are likely to be especially relevant for online interactions. Consequently, it is crucial to understand how and why people judge faces from photographs in this way, in order to build a solid theoretical base with which to understand these important real-life implications.

1.3. How accurate are facial impressions?

An important second question is the extent to which these facial first impressions are accurate. However, the evidence for whether first impressions from face photographs are accurate is currently somewhat mixed (see Rule, Krendl, Ivcevic, & Ambady, 2013; Todorov et al., 2015 for recent reviews). First, a number of studies have focused on impressions of the Big Five personality factors of extraversion, agreeableness, openness to experience/intellect, conscientiousness and neuroticism from faces. In these studies, accuracy is defined as perceivers making social judgements from faces that match the targets' self-reports, and these studies are also often characterised by high control over the stimuli images (Kramer & Ward, 2010, 2011; Little & Perrett, 2007; Penton-Voak, Pound, Little, & Perrett, 2006; Rule et al., 2013). This research has shown that some personality inferences are judged with above-chance accuracy from faces; especially judgements of extraversion, and often also agreeableness and neuroticism (Jones, Kramer, & Ward, 2012; Kramer & Ward, 2010, 2011; Little & Perrett, 2007; Penton-Voak et al., 2006; Rule et al., 2013). There may therefore be a "kernel of truth" (Penton-Voak et al., 2006) in the idea that we can accurately judge others' Big Five personality from their face, at least with carefully controlled face stimuli.

Another line of research has focused on social group membership, and this has found that people are able to use facial photographs to accurately judge sexual orientation

(Rule, Ambady, & Hallett, 2009; Rule, Ishii, Ambady, Rosen, & Hallett, 2011) religious belief (Rule, Garrett, Ambady, & Rustichini, 2010) and political affiliation (Rule & Ambady, 2010). Again, accuracy here refers to above-chance identification of the targets' self-reported sexual, religious or political orientation. However, Olivola and Todorov (2010b) have shown that participants' facial judgements of political affiliation and sexual orientation, among other social characteristics, are less accurate than judgements would be if perceivers relied simply on base-rate information about the distribution of these characteristics in the population.

A number of studies have now examined the accuracy of trustworthiness judgements in particular; however, the evidence for accurate moral character assessment from facial inferences is relatively weak (see Rule et al., 2013 for a review). On the one hand, some studies have found slight evidence that trustworthiness judgements are accurate. In a classic study, Bond and colleagues (C. F. Bond, Berry, & Omar, 1994) found that people who were perceived as facially dishonest were more willing to sign up for a study involving deceptive behaviour (yet note that Masip & Garrido, 2001 failed to replicate this finding). More recently, Stirrat and Perrett (2010) found that the ratio of male faces' facial width to height (fWHR) correlated with the proportion of trustworthy decisions made in economic games (i.e. fairly dividing money with a partner; Spearman's r = -.40), but this was not found for female faces. Greater male target fWHR was in turn linked to perceptions of untrustworthiness (Stirrat & Perrett, 2010). Porter and colleagues (Porter, England, Juodis, Ten Brinke, & Wilson, 2008) also found that the faces of Nobel Peace Prize winners were judged as trustworthy at above chance levels with a long (but not short) presentation time at both an initial and a repeated viewing, but that the faces of criminals were not generally accurately judged as untrustworthy. They concluded that the evidence for facial trustworthiness being valid was therefore rather weak (Porter et al., 2008). Finally, one other study found that photographs of violent criminals were judged as looking slightly more violent than photographs of non-violent criminals (Stillman, Maner, & Baumeister, 2010). This was mediated by presumably more veridical cues to perceived facial age and masculinity.

On the other hand, the majority of studies have found that people are not accurate when perceiving the trustworthiness of targets from facial photographs (Bonnefon, Hopfensitz, & De Neys, 2013; Efferson & Vogt, 2013; Rule et al., 2013; Zebrowitz, Voinescu, &

Collins, 1996). In these studies, target 'trustworthiness' was measured by asking clinicians' opinions of the targets (Zebrowitz et al., 1996), examining targets' cheating behaviour on a test or comparing photographs of criminals and non-criminals (Rule et al., 2013); or measuring targets' behaviour in economic trust games (Bonnefon et al., 2013; Efferson & Vogt, 2013; Rule et al., 2013). Interestingly, one study found that the accuracy of trustworthiness judgements depends on the target photograph, so that the trustworthiness of decisions during a game were discriminated at above chance (around 63%) from photographs of people taken at the moment of their decision, while photographs of the same targets taken before the game could not be distinguished on their trustworthiness above chance (Verplaetse, Vanneste, & Braeckman, 2007).

Lastly, a number of studies have examined the accuracy of judgements of intelligence from face photographs. Kleisner and colleagues found that observers' ratings of perceived facial intelligence accurately predicted intelligence as measured with an IQ test for male but not female faces (Kleisner, Chvátalová, & Flegr, 2014). Zebrowitz and colleagues carried out a meta analysis on studies examining the accuracy of perceived facial intelligence, and found a small but significant correlation of around .3 between intelligence measured with IQ tests and the average of strangers' ratings across studies (Zebrowitz, Hall, Murphy, & Rhodes, 2002; see also Rule et al., 2013). In their own study of intelligence across the lifespan, they also found that the relationship between perceived and actual intelligence was especially strong in childhood (Zebrowitz et al., 2002). Attractiveness significantly mediated the relationship between perceived and actual intelligence through multiple biological and environmental pathways, including 'good' genes (especially in childhood), a link with socio-economic status and self-fulfilling prophecies (Zebrowitz et al., 2002). In a later study they replicated their finding of accuracy between measured and perceived facial intelligence, but only for targets below the median in attractiveness, suggesting that participants were sensitive to low fitness or 'bad' genes (Zebrowitz & Rhodes, 2004). Lastly, the research showing that facial judgements of competence and dominance predict job performance, is also somewhat suggestive of accuracy (Rule & Ambady, 2008, 2009; Walker Naylor, 2007). However, these results could also represent self-fulfilling prophecy effects (i.e. perhaps companies hire more competent looking CEOs).

Finally, it is worth highlighting that measuring the accuracy of facial impressions is bound up with issues of stimuli selection and measurement choice. First, perceivers' accuracy may also depend on the sample of face photographs used, since there is often as much variability in judgements of different photographs of the same individual as in judgements of photographs of different individuals (Jenkins, White, Van Montfort, & Burton, 2011; Todorov & Porter, 2014; this will be returned to in more depth in the **Ambient images section)**. Second, accuracy can be measured in (at least) two ways: by correlating an external criterion with either the average rating across individual raters or with each individual rater's judgement. The first measurement naturally gives a higher estimate of accuracy than the second measurement, since averaging tends to remove noise in the judgements of individual raters. In other words, this first measurement does not imply that each individual rater is necessarily especially accurate but instead measures the accuracy of the group consensus (Todorov, Said, et al., 2011). The choice of external criterion may of course also affect the apparent accuracy of the judgements. Moreover, Rule and colleagues suggest that it is not yet clear whether trustworthiness in particular is a unitary trait since it is possible that people who are trustworthy in one situation may not be in other situations (Rule et al., 2013). This possibility remains to be empirically tested.

In summary, it seems that evidence is mixed at best for the accuracy of perceivers' facial impressions. There is probably a small kernel of truth to some judgements, especially for more visible or easily signalled attributes such as extraversion and group membership. For character judgements, it seems that perceivers are not generally accurate, at least when judging trustworthiness from a single facial photograph of that individual. Moreover, accuracy is usually measured as correct categorisation at rates only slightly higher than chance, which is "a rather feeble benchmark" as Todorov and colleagues point out (Todorov et al., 2015, p15.14), especially as facial judgements are often less accurate than judgements made simply on knowledge of the base-rates of characteristics in the population (Olivola & Todorov, 2010b). For this thesis however, it is immaterial whether facial social impressions are accurate, since, as outlined earlier, perceivers rapidly and spontaneously make these impressions, and these impressions have important real-life consequences. These findings make it interesting to understand facial impressions, regardless of their accuracy. In addition, as will be seen in the next

section, many facial cues to impressions are based on category groups. In this sense, facial impressions can be thought of in much the same way as any other stereotype, where accuracy is usually not a defining feature.

1.4. Face perception approaches to first impressions

1.4.1. Cues to facial impressions

There is a large and sometimes disparate face perception literature, which mainly attempts to examine which facial cues subserve character impressions. Social information that we readily retrieve from the face includes perceived age, sex, race, physical attractiveness and emotional expression (Bruce & Young, 2012; Montepare & Opeyo, 2002; Oosterhof & Todorov, 2008) and these have all been outlined as cues to character judgements. This review will only provide a brief overview of these main cues and how they relate to social impressions from faces.

1.4.1.1. Facial age

We have a good understanding of age-related changes in the face (Bruce & Young, 2012). Researchers have developed mathematical ('cardioidal strain') transform functions that closely track real and perceived age-rated facial structural growth, possibly because they seem to index gravitational forces on the bone growth of the young head (Mark et al., 1980; Pittenger & Shaw, 1975). While these changes in the head shape are used by perceivers to gauge the age of younger targets, changes in skin texture are more informative when assessing the age of older adults (above 30: Mark et al., 1980). These include increased wrinkles, skin and hair greying, and less colour evenness with increased age (Burt & Perrett, 1995; Gunn et al., 2009; Mark et al., 1980). Moreover, perceived age is a good biomarker of 'biological age' (i.e. how healthily a person is aging) with participants relying on the aforementioned cues (Gunn et al., 2009).

In terms of social characteristics, the faces of elderly people are perceived as less attractive, likeable, sociable, warm; less strong, healthy and active; and more emotional (Ebner, 2008; Zebrowitz, Fellous, Mignault, & Andreoletti, 2003). For a sample of the faces of young adults (with an initial mean age of around 21 years), increasing the

apparent age of these faces by a few years increased their perceived dominance and ambition (Boothroyd, Jones, Burt, & Perrett, 2007). In a related fashion, the more childlike, or 'babyfaced' the target looks (that is, showing a lack of facial maturity, which is conceptually distinct from perceived age), the more lovable, kind, socially submissive and less threatening they also seem (D. S. Berry & McArthur, 1986; Zebrowitz, 2005). Large eye size, thinner eyebrows and rounder faces seem to subserve these babyfacedness impressions (Zebrowitz & Montepare, 1992); with adult faces' objective physical resemblance to real babies' faces predicting these social impressions (Zebrowitz et al., 2003).

1.4.1.2. Facial sex

We are highly accurate at judging someone's biological sex from their face, with 96% correct classification in one study (Bruce et al., 1993). Compared to female faces, male faces typically have larger noses, jaws and chins; smaller eyes, more prominent brows and less full cheeks (E. Brown & Perrett, 1993; Bruce et al., 1993; Burton, Bruce, & Dench, 1993). In terms of texture, female faces typically have less facial hair (Burton et al., 1993); and show more contrast between (pale) skin and (dark) eyes/lips (R. Russell, 2009). Lips with redder hue also look more feminine (Stephen & McKeegan, 2010). Perceivers seem to use many of these cues together to categorise faces on sex (Burton et al., 1993; Dupuis-Roy, Fortin, Fiset, & Gosselin, 2009).

Female faces, and more feminine-looking or less masculine-looking faces of both sexes are perceived as less dominant and more socially orientated (Hess, Adams Jr, & Kleck, 2005; Perrett et al., 1998). In part, these social impressions can be explained by female faces also looking more babyfaced (Friedman & Zebrowitz, 1992) and also because female faces objectively resemble happiness expressions, and male faces, anger (Adams Jr, Hess, & Kleck, 2015; Becker, Kenrick, Neuberg, Blackwell, & Smith, 2007; Hess, Adams Jr, & Kleck, 2004). Large jaws, indicating increased facial masculinity, also subserve impressions of dominance (Keating, Mazur, & Segall, 1981). Finally, feminine-looking female faces are almost always more attractive, but both feminine-looking and masculine-looking male faces can be attractive (see the **Facial attractiveness** section).

1.4.1.3. Facial race

While the concept of 'race' is debated and may have no biological reality (Zebrowitz, Kikuchi, & Fellous, 2010), it is clear that people can easily and quickly discriminate faces based on their apparent racial group (Montepare & Opeyo, 2002; Rossion & Michel, 2011). Although there are clearly many possible racial and cultural groups, most face research has focused on three racial categories: 'Asian', 'White' and 'Black' faces. 'Asian' faces typically have the narrowest eyes, look most babyfaced and, even when neutrally posed, have an objective structural resemblance to surprise expressions, as determined by a neural network (Strom, Zebrowitz, Zhang, Bronstad, & Lee, 2012; Zebrowitz et al., 2010). 'Black' faces typically have the darkest skin, relatively wide noses, and an objective resemblance to happiness expressions and masculine features; while 'White' faces typically have the lightest skin, more varied eye and hair colour, more convex faces with 'pointy' noses, and an objectively feminine appearance (Hill et al., 1995; Johnson, Freeman, & Pauker, 2012; Rossion & Michel, 2011; Strom, Zebrowitz, Zhang, Bronstad, & Lee, 2012; Zebrowitz et al., 2010).

Perceivers use both textural and shape cues in categorising faces on race, with slightly more reliance on shape cues, at least for 'White' perceivers (Hill et al., 1995; Strom et al., 2012). Recently, studies have shown that perceivers can also discriminate between faces that are more or less phenotypic for their ethnicity, within an ethnic category (Blair, Judd, Sadler, & Jenkins, 2002; Maddox & Gray Chase, 2004). For example, some faces, while appearing 'White', may have facial features which, to a greater or lesser degree resemble 'Black' faces, such as a relatively wider nose or darker skin tone (Blair et al., 2002; Maddox & Gray Chase, 2004).

In terms of social attributions, 'Black' faces and more phenotypically 'African' faces are perceived as more athletic, less reserved and more masculine; while 'Asian' and more phenotypically 'Asian' faces are seen as more reserved, less dangerous and more competent (Strom, Zebrowitz, Zhang, Bronstad, & Lee, unpublished manuscript; Zebrowitz, Bronstad, & Lee, 2007; Zebrowitz et al., 2010). 'White' and more phenotypically 'White' faces are generally perceived somewhere in between (Johnson et al., 2012; Zebrowitz, Bronstad, & Lee, 2007; Zebrowitz et al., 2010). In part, these inferences seem to be influenced by a facial structural overlap with emotional

expression and/or sexually dimorphic features, as assessed by objective facial measurements (Johnson et al., 2012; Zebrowitz et al., 2010).

1.4.1.4. Facial attractiveness

It is now understood that the perception of facial beauty is substantially culturally universal (Cunningham, Roberts, Barbee, Druen, & Wu, 1995; Thornhill & Gangestad, 1999) and that attractiveness perceptions are driven in part by evolution via sexual selection (Little, Jones, & DeBruine, 2011; Rhodes, 2006). While cues to attractiveness perceptions are less well specified than, for example, age cues, we nevertheless have a good understanding of some of the cues underlying facial beauty.

Galton (1879) first showed that the average of a set of faces is more attractive than the individual faces entering the average (see also Langlois & Roggman, 1990; and Rhodes 2006 for a meta-analysis). Moreover, facial ordinariness and attractiveness are correlated (Light, Hollander, & Kayra-Stuart, 1981). Over and above averageness, facial symmetry has also been recognised as being attractive (Thornhill & Gangestad, 1999) and has also been linked to indicators of target health (although the evidence for this is variable: Little et al., 2011). As described in the **Facial sex** section, a robust literature links feminine features with perceived attractiveness for females (Rhodes, 2006; R. Russell, 2010) although the relationship between femininity and healthiness/good genes is somewhat unclear (Little et al., 2011). For male faces, the picture is more complicated since both femininity and masculinity can be perceived as attractive, depending on the stimuli used, whether face shape and/or reflectance is manipulated, the female perceivers' reproductive cycle and fertility, the perceivers' preference for masculinity and the healthiness of the immediate and societal context (see Little et al., 2011, and Rhodes, 2006 for reviews). Finally, facial texture cues to health are also attractive. For example, yellow (carotenoid) skin colouring, which is linked to fruit and vegetable consumption, is perceived as attractive (Lefevre, Ewbank, Calder, von dem Hagen, & Perrett, 2013; Stephen, Law Smith, Stirrat, & Perrett, 2009). An even skin tone is also perceived as attractive and healthy (Fink, Grammer, & Thornhill, 2001). Youthfulness cues are also attractive for female faces, probably in part again due to associations with fertility and health (Buss & Schmitt, 1993; Thornhill & Gangestad, 1999).

There is substantial evidence regarding the importance of facial attractiveness in character impressions (Dion, Berscheid, & Walster, 1972; Little et al., 2011). Indeed, the *'what-is-beautiful-is-good'* effect was coined to describe the finding that physically attractive faces are ascribed other positive attributes such as sociability, popularity and intellectual competence (Dion et al., 1972; Eagly, Ashmore, Makhijani, & Longo, 1991). A more recent study has linked physical attractiveness stereotyping to anomalous overgeneralisation effects, whereby faces perceived as below median attractiveness are indeed accurately perceived in intelligence (r = .25 between perceived intelligence and IQ scores) but this link between attractiveness and intelligence is then inaccurately overgeneralised to attractive faces (i.e. '*what-is-unattractive-is-bad*: Zebrowitz & Rhodes, 2004).

1.4.1.5. Emotional expression

The previous sections concentrated more on long-lasting (structural and textural) features of a face; however, faces are clearly highly dynamic and very emotionally expressive. We have a good understanding of the distinctive muscle patterns that contribute towards particular expressions, as Ekman and colleagues have systematically catalogued the possible muscle combinations of the face that result in perceptual changes in the Facial Action Coding system (see Cohn, Ambadar, & Ekman, 2007 for a recent guide). Certain muscle patterns are also reliably recognised as reflecting particular emotions across cultures, supporting the idea that these 'basic' emotional expressions have a long evolutionary history (Ekman, 1972). However, there are also clearly emotional cultural 'dialects', so that different events elicit these emotions and different rules govern when and to what extent emotional expressions are displayed across culture (Ekman, 1972, 1994; see Elfenbein & Ambady, 2002 for a meta-analysis). Most studies include happiness, sadness, anger, disgust and fear as basic emotions, with several others currently being debated, including contempt, pride and surprise (Elfenbein & Ambady, 2002; Tracy & Randles, 2011).

Emotional expressions are clearly linked to character attributions: for example, people who appear to be angry also look less affiliative and dominant; while people who appear to be happy instead look affiliative but submissive (Montepare & Dobish, 2003;

Oosterhof & Todorov, 2008; Zebrowitz, Kikuchi, & Fellous, 2007). Surprise and fear impressions meanwhile, lead to lower dominance; and surprise also leads to higher affiliation (Montepare & Dobish, 2003; Zebrowitz, Kikuchi, et al., 2007).

Secord (1958) has suggested that character inferences from emotions occur through *temporal extension* whereby people, upon perceiving a transient emotional expression on a target's face, misattribute the temporary cue to a fixed disposition. Zebrowitz and colleagues have extended this argument into *emotional overgeneralisation*, where this dispositional misattribution can also happen even when target faces are actually expressionless, but merely structurally resemble emotional expressions (D. S. Berry & McArthur, 1986). In support of these mechanisms, the expressionless faces of people perceived by raters as threatening were misclassified as 'angry' by a neural classifier which had been trained on emotional and neutral faces (among other trait-emotion links: Said, Sebe, & Todorov, 2009). Certain emotional expression cues do indeed overlap with cues to sex (Adams Jr, Nelson, Soto, Hess, & Kleck, 2012; Hess et al., 2004; see the **Facial sex** section) and race (Zebrowitz et al., 2010b; see the **Facial race** section).

Finally, there are other, specific cues in the face that have been linked to character traits: for example, direct gaze increases dominance perceptions relative to averted gaze (Main, Jones, DeBruine, & Little, 2009). Carré, McCormick, and Mondloch (2009) have shown that the ratio of facial width to facial height (fWHR), predicts targets' perceived and actual aggression (Carré, Morrissey, Mondloch, & McCormick, 2010; Geniole, Molnar, Carré, & McCormick, 2014) although other studies find that increasing fWHR decreases perceptions of trustworthiness (Stirrat & Perrett, 2010) and increases perceptions of dominance (Alrajih & Ward, 2014) instead. Lastly, others have investigated cues to more ambiguous perceived group membership; for instance, femininity in male faces and masculinity in female faces predicts perceptions of homosexuality (Freeman, Johnson, Ambady, & Rule, 2010); while perceptions of social class, attractiveness, intelligence and power predict perception of political membership (Bull & Hawkes, 1982; Bull, Jenkins, & Stevens, 1983; Rule & Ambady, 2010).

1.4.1.6. Modelling social impressions from faces

From the previous sections, it should be clear that there is considerable overlap between the various social judgements made from faces and the cues that subserve these judgements. Without an overall structure to the judgements made or cues underlying these judgements, it is difficult to see the overall picture underlying these results or to know which judgements to focus on in future research. To address this, Oosterhof and Todorov (2008) used principal components analysis to combine multiple social judgements into one overall model. This approach allows future research to focus on these key dimensions of social judgement, thereby setting out an important and useful theoretical framework for the field.

In this approach, Oosterhof and Todorov (2008) first asked participants for their spontaneous impressions of a set of sixty-six standardised facial photographs. They then categorised these spontaneous impressions by how frequently they were mentioned by participants, and thus reduced the initial impressions to a tractable set of fourteen traits. Dominance was also included, to mirror previous models of interpersonal trait evaluation (Wiggins, 1979). Oosterhof and Todorov (2008) then collected new quantitative ratings of these fifteen traits using the same faces, and used principal components analysis (PCA) to reduce the thirteen most reliable trait judgements into underlying dimensions of variance. This resulted in two dimensions, closely approximated by trustworthiness and dominance judgements, which together explained around 82% of the variance in the initial ratings (see **figure 1.1.**).

The trustworthiness dimension involves the perceived *intentions* of the target ('Do they want to hurt or help me?' Oosterhof & Todorov, 2008). It is based on an emotion overgeneralisation; so that faces which structurally resemble happiness are viewed as approachable and thus perceived as trustworthy, while faces which structurally resemble anger are to be avoided, and thus perceived as untrustworthy (Todorov, 2008; Zebrowitz et al., 2010). The dominance dimension is based on the perceived *capability* of the target ('Can they hurt or help me?') and underlying this dimension are facial cues to physical strength, masculinity and maturity (Fink, Neave, & Seydel, 2007; Oosterhof & Todorov, 2008).

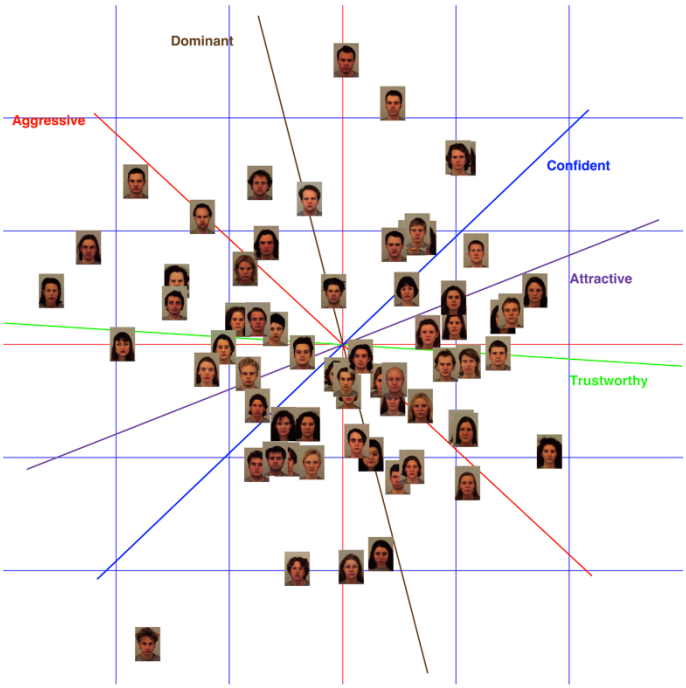

Figure 1.1. A two-dimensional plot of emotionally neutral faces, plotted as a function of their loading on the two dimensions of valence/trustworthiness (horizontal orange line) and dominance (vertical orange line). The other coloured vectors depict selected facial ratings in this space and units are standard deviations. Reproduced from Oosterhof and Todorov (2008).

It should be noted that Oosterhof and Todorov (2008) state that the finding of these two dimensions and their interpretation remained unchanged without the dominance ratings that were added in by the authors. Finally, it is also worth pointing out that Oosterhof and Todorov (2008)'s work was characterised by a careful control of the stimuli; with initial descriptions and ratings made to standardised face photographs, and with later studies replicating these results with highly controlled computer generated ("FaceGen") faces. Moreover, two very similar dimensions were found in a conceptual replication by an independent research group (Walker & Vetter, 2009).

Oosterhof and Todorov (2008) argue that these dimensions are fundamental to facial first impressions because together they represent the appraisal of a target's threat. They also invoke an evolutionary account of these dimensions, which situates them in our primate ancestry (Todorov, Said, Engell, & Oosterhof, 2008). This itself implies that these dimensions should be universal: that is, these dimensions should appear across multiple cultures and social groups. The universality idea is interesting and potentially

influential, but has yet to be directly tested with a non-Western model of facial first impressions.

Earlier, this review concluded that evidence for the accuracy of trustworthiness judgements is weak, which might seem to contradict the claim that these judgements represent a threat appraisal mechanism rooted in our primate ancestry. One may ask why natural selection would have equipped us with threat appraisal mechanisms that are inaccurate. One reason could be that it might be better for perceivers to be highly sensitive to the negative end of these judgements, at the expense of accuracy *per se*. Indeed, Oosterhof and Todorov (2008) found evidence that this was the case. In addition, it might be better for perceivers to be especially sensitive to impressions of threat as they occur in the moment, at the expense of being accurate about stable dispositions. This echoes the emotional extension and temporal overgeneralisation hypotheses discussed previously (see also the **Discussion**).

In summary, this theoretical approach therefore aligns with the findings reviewed in the previous section, which implicate "bottom up" cues to trait impressions; that is, cues which are present in the visual stimulus of the face itself. This review will next briefly outline a new approach to face sampling ('ambient images') then discuss alternative, 'top-down' routes to generating first impressions.

1.4.2. The 'ambient image' face approach

Recently, Jenkins, Burton and colleagues have argued for the use of more variable face photographs in face perception studies (Burton, Jenkins, & Schweinberger, 2011; Jenkins et al., 2011). They point out that many face perception studies use a single, or perhaps only a couple of photographs of an individual. These photographs are also often standardised to be highly alike in terms of ambient environmental characteristics (for example, with uniform lighting, frontal facing, greyscale and so forth). Yet, as Jenkins, Burton and colleagues explain, naturalistic photographs of individuals, which they term "ambient images", differ on a wide variety of cues, and these cues may change impressions or recognition across photographs containing the same person (Burton et al., 2011; Jenkins et al., 2011). For instance, real-life photographs of individuals' faces (see **figure 1.2.**) often differ on their facial pose, head direction, emotional expression,

as well as environmental characteristics such as the lighting, the camera angle and camera type, the extra-facial background, and facial paraphernalia such as glasses, facial hair and piercings.

This idea of image variability is important not just for identity recognition, but also for social impressions of faces from photographs. For example, using ambient images which were sampled from the internet, Jenkins and colleagues (2011) showed that impressions of individuals' facial attractiveness differed more depending on the photograph chosen than the individual actually depicted. That is, the within-person variability was greater than the between-person variability. Since then, Todorov and Porter (2014) have shown that the within-person variability in ambient images was also either equivalent to or exceeded the between-person variability for facial judgements of competence, creativity, cunning, extraversion, meanness, smartness and trustworthiness. This suggests that the variability in naturalistic photographs can be substantial and that this variability is itself important in making impressions. By controlling these characteristics, researchers risk ignoring or altering the very cues that perceivers use to make first impressions.

Figure 1.2. Example ambient face images. Reproduced from Vernon, Sutherland, Young and Hartley (2014).

The idea of using ambient images to broaden the scope of cues available to perceivers and to allow for a more naturalistic examination of facial impressions, has important implications for the facial impressions model of trustworthiness and dominance described previously (Oosterhof & Todorov, 2008). Indeed, the ambient image approach raises the question of whether this two-dimensional model, which was built on carefully standardised images (Oosterhof & Todorov, 2008; see also Walker & Vetter, 2009), would still exist for impressions made to a set of naturalistic, highly varying face images. This is an issue at the heart of this approach, since basing future research around a dimensional model only has utility if the model captures the main cues

involved. This idea will be returned to at the end of the introduction, in **Issues investigated** section.

Finally, the finding that different photographs of the same individual can be perceived differently in terms of attractiveness (Jenkins et al., 2011) or social impressions (Todorov & Porter, 2014) highlights that the results of most face perception studies pertain to single photographs of individuals, rather than that individual identity *per se*. Indeed, there may be no absolute "trustworthiness" rating one can give to a single individual, if this judgement depends on the photograph chosen (see also the **Accuracy of facial impressions** section). Where this research in this thesis describes 'facial trustworthiness' therefore, this should be taken to mean the trustworthiness of a single face photograph. This thesis also uses the term 'ambient face images' only to refer to the use of highly, naturalistically varying facial photographs in order to examine a wider range of potential cues to facial first impressions, not the use of multiple photographs of the same individuals. In future, it may be interesting to explore this distinction further with regards to social impressions, in order to separate identity-specific cues from changeable cues (see **Chapter 6, Future directions** section).

1.4.3. Interim summary

In summary, this section has reviewed cues to social impressions, with the general argument that perceivers over-generalise relatively veridical cues present in faces (e.g. genuinely smiling faces) to enduring trait characteristics (e.g. trustworthiness inferences: Zebrowitz et al., 2003). Recently, an examination of the structure of social inferences from faces has narrowed down the multiple possible cues and social inferences into two key dimensions, trustworthiness and dominance (Oosterhof & Todorov, 2008). Finally, more recently still, researchers have argued for the importance of using highly variable ambient image faces in order to fully sample cues available to perceivers (Burton et al., 2011; Jenkins et al., 2011). In the next section, top-down routes to impressions will be considered.

1.5. Social psychological approaches to impressions

1.5.1. Stereotypes

> For the most part we do not first see, and then define, we define first and then see. In the great blooming, buzzing confusion of the outer world we pick out what our culture has already defined for us, and we tend to perceive that which we have picked out in the form stereotyped for us by our culture.

(Lippmann, 1922), *Chapter VI: Stereotypes. p. 81.*

The word 'stereotype' (literally, 'fixed impression'), was first used in the social sense by Lippmann (1922), who likened them to 'pictures in the head'. In current psychological understanding, stereotypes have come to mean knowledge associated with a category group, including both descriptive and prescriptive/proscriptive beliefs (Hilton & Von Hippel, 1996; Prentice & Carranza, 2002). These beliefs can encompass typical roles, behaviour (Kunda, Sinclair, & Griffin, 1997), occupations, physical appearances, and personality traits (Deaux & Lewis, 1984; Six & Eckes, 1991). Stereotyping, by referring to the knowledge associated with category groups is therefore distinct from categorisation, which refers to the perception of individuals as belonging to social groups (Lepore & Brown, 1997), and prejudice, which is the evaluation of a social group (as opposed to a cognitive belief: Amodio & Devine, 2006).

Lippman (1922) considered stereotyping to be functional because stereotypes help busy perceivers quickly make sense of their complex world, providing efficiency at the expense of accuracy (the *cognitive miser* account: see Allport, 1979; Fiske & Taylor, 1984). Later accounts have instead emphasised the motivated nature of stereotyping and highlighted the flexibility of the perceiver, who uses stereotypes as only one cognitive tool alongside more considered cognitive appraisal (the *'motivated tactician'* metaphor: see Fiske & Taylor, 1991; Macrae & Bodenhausen, 2000). Spears and Haslam (1997) have also emphasised stereotyping as adaptive and motivated, and term the perceiver a *'meaning seeker'* who rapidly constructs often quite complex stereotypes to fill in the gaps in their direct perception. Stereotyping therefore has the potential to facilitate social interaction, by allowing one to make an extensive impression of an individual even in the first moments of meeting.

In relation to our main topic of first impressions, stereotypes essentially are first impressions of individuals made on the basis of abstract knowledge about their social or cultural group. There is a rich social psychological literature that implicates stereotyping in the formation of trait impressions entirely outwith the face (see Hilton & Von Hippel, 1996 for a review). Indeed, perceivers make quite extensive judgements of targets by extrapolating information from their category group (Hilton & Von Hippel, 1996). For example, gender stereotypes may allow perceivers to predict that in general, males will be agentic and dominant, and females will be warm and caring; and to evaluate that these are desirable characteristics for men and women to have respectively (Prentice & Carranza, 2002). One of the most prominent and widely used models of the content of stereotypes in relation to social impressions will be outlined next.

1.5.2. The stereotype content model of impressions of groups

Over the last decade, Fiske and colleagues have outlined the *'stereotype content model'* (SCM), which has become prominently used in social psychology to explain the main dimensions of stereotype content. The SCM has shown that two dimensions subserve impressions of a wide variety of social groups (for example, 'the homeless', 'feminists' and 'Asians').

The first dimension is described as 'warmth' and describes groups that perceivers tend to either like or dislike (Cuddy, Fiske, & Glick, 2008; Fiske, Cuddy, & Glick, 2007; Fiske, Cuddy, Glick, & Xu, 2002). This dimension relates to a judgement of the intentionality of the target group, so that groups that are perceived to have benevolent intentions (such as 'the elderly') tend to be perceived as being warm and likable, while groups with apparently competitive intentions or who challenge the status quo (such as 'feminists') tend to be seen as cold and disliked. The second dimension is described as 'competence' and describes groups that tend to be either respected or disrespected (Cuddy et al., 2008; Fiske et al., 2007). This dimension relates to a judgement of perceived capability, so that groups that are perceived to be capable of carrying out their intentions (good or bad) tend to be respected and seen as competent (such as 'the rich') while those who are less capable tend not to be respected, nor seen as competent (such as 'the poor').

Interestingly, two-dimensional models of impressions of individuals, groups, and the self, appear across social psychology, with apparently similar dimensions to warmth and competence. For example, communality and agency dimensions appear for judgements of the self (Abele & Wojciszke, 2007; Wojciszke, 2005), social desirability and intellectual dimensions appear in person perception impressions (Rosenberg, Nelson, & Vivekananthan, 1968) and an early model of person perception from verbal labels, the interpersonal circumplex model, can be characterised as having affiliation and dominance dimensions (Wiggins, 1979). Each model shares the central idea of a division between evaluative or social traits on the one hand, and agentic or competence based traits on the other.

Strikingly, the SCM and similar models also bear an apparently strong resemblance to the two-dimensional model of facial impressions found by Oosterhof and Todorov (2008). The SCM dimension of warmth appears highly similar to the description of the facial dimension of trustworthiness, and the SCM dimension of competence seems highly similar to the description of the facial dimension of dominance (see **figure 1.3.**). Moreover, both models fundamentally link the first dimension to a judgement of intentionality and the second dimension to capability; both models have an underlying theoretical analysis that argues that social impressions are evolutionary adaptations, and both models either explicitly or by implication assume that these judgements are universal across culture. The stereotype content model has found evidence for the universality of the warmth and competence dimensions (Cuddy et al., 2009), but this has yet to be tested for the facial dimensions. This point will be returned to in the **Issues investigated** section.

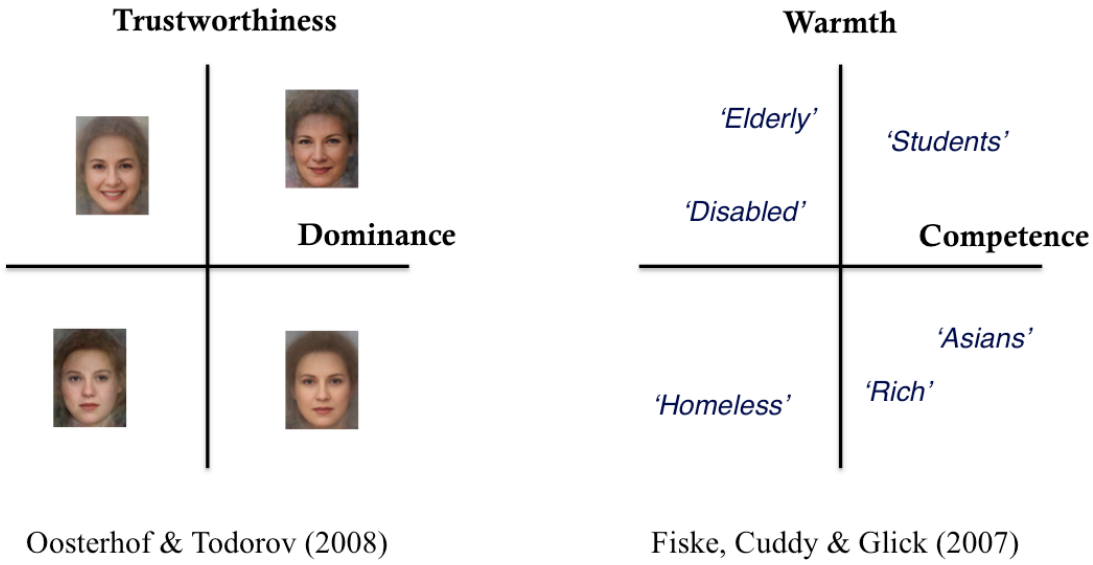

Figure 1.3. A schematic diagram representing Oosterhof and Todorov's (2008) leading model of facial first impressions and Fiske et al. (2007)'s leading model of group stereotypes.

Taken together, this analysis of the social and facial models suggests that the same dimensions may subserve impressions of faces (which could rest on 'bottom up' visual cues) and stereotypes of groups (which can only be 'top down', since only abstract verbal labels are given). If this is correct, it suggests that these dimensions represent fundamental aspects of human social perception, with the same key judgements being made when first seeing an individual face and when thinking about abstract groups. This would be an elegant finding, spanning multiple fields and very different approaches to examining social perception. However, the similarity of warmth and trustworthiness judgements, or competence and dominance judgements, has not been yet directly quantified for facial impressions. This theme will be returned to in the **Issues investigated** section, after a discussion of studies that examine the overlap between categorisation and associated stereotypes with face perception.

1.5.3. Social groups, stereotyping and face perception

In an insightful investigation into facial first impressions, Hassin and Trope (2000) carefully emphasised that both facial cues (*'reading from faces'*) and perceivers' expectations of individuals' characteristics *('reading into faces'*) are important in creating a first impression. The previous sections of this review illustrate reading *from*

faces. As an example of reading *into* faces, Hassin and Trope (2000) ask if we would perceive Einstein as having such a large forehead if we did not know of his intelligence beforehand. They further speculated that these two 'reading' processes recursively interact to produce the final person impression.

From prior expectations based on knowledge about individuals, it is a short jump to argue that prior knowledge based on cultural groups (stereotyping) is also important in generating first impressions of faces. Secord (1958) also suggested that such a mechanism exists, which he called *categorisation* (alongside his mechanism of temporal extension which was outlined previously). Indeed, Secord's ideas were strikingly prescient in outlining many of the areas of research into facial first impressions today. Arguably, since then, face perception research has tended to overlook insights from social psychology (Quinn & Macrae, 2011).

More recently, research has again started examining how perceivers categorise faces into different groups, along with associated abstract knowledge (i.e. stereotypes: Quinn & Macrae, 2011). Much of this work examines how the categorisation of a face as belonging to one group (for example, as masculine or feminine), then affects the categorisation of a face into other groups, either simultaneously or downstream (for example, as heterosexual or homosexual: Freeman et al., 2010). For instance, Freeman and Ambady (2011) recently outlined a model of person construal that explicitly also includes a role for top-down stereotyping on the categorisation of face stimuli (see **figure 1.4.**). Visual input from faces (among other cues) feeds into categories (such as 'age' or 'gender'), which feed into associated stereotypes (such as 'aggressive'). Higher order states (such as the task or individual motivations) affect this process by directing attention. Feedback and lateral interactions are also included in this model. Their model is a recurrent connective network, is grounded in neural dynamics, and emphases the interactivity of prior knowledge with current perceptual input. It builds on earlier work by Kunda and Thagard (1996); and Read and Miller, (1998). For present purposes, it again highlights the point that stereotypes can be held abstractly and yet can also be reached through the immediate visual input of a face, and it posits bi-directional links between the two sources of information.

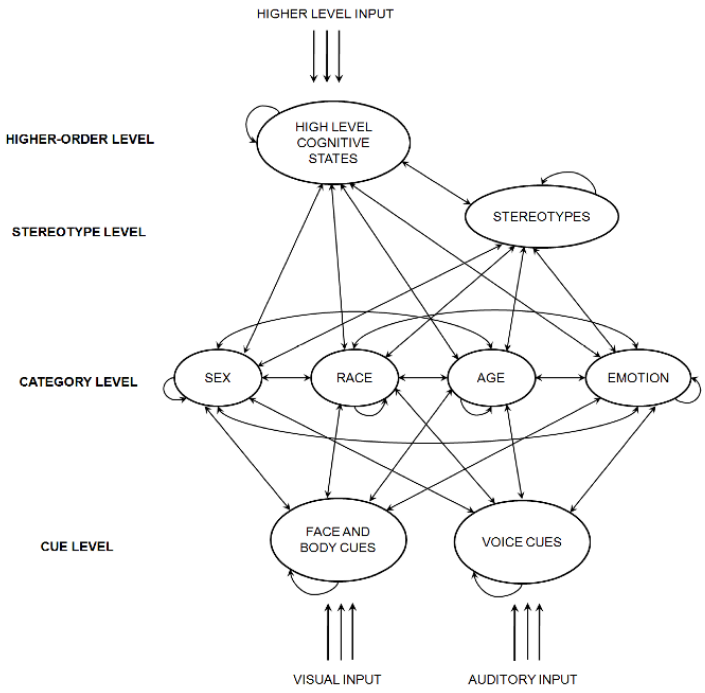

Figure 1.4. The Dynamic Interactive Model of person construal. Reproduced from Freeman and Ambady (2011).

Empirical findings increasingly support these integrative, dynamic models. For example, Johnson, Freeman, and Pauker (2012) showed that face sex categorisation is also affected by the race of the face, so that female faces were categorised faster if they were also Asian relative to White, and White relative to Black (vice versa for male faces). This bias in categorisation was also mediated by individual differences in strength of implicit associations between gender and race, as measured by the implicit association test. In another recent study, faces were more likely to be categorised as 'Black' rather than 'White' if they wore low status clothes compared to high status clothes (Freeman, Penner, Saperstein, Scheutz, & Ambady, 2011). This tendency was especially strong for racially ambiguous faces. Freeman and colleagues (2011) argued that this finding is due to stereotypic associations between status and race affecting the visual racial categorisation task. Bijlstra and colleagues also invoked race/sex stereotyping to explain reaction time biases between faces varying on sex or race in emotion categorisation tasks (Bijlstra, Holland, & Wigboldus, 2010). Interestingly, Dzhelyova and colleagues (2011) also found facilitation effects in reaction times and accuracy for categorising female faces as trustworthy (or vice versa, trustworthy faces

as female), which they pointed out could be due to the stereotype of women as trustworthy.

Straightforward stereotyping effects linked to face categories are not always found. In order to investigate the effect of stereotyping on the overlap between emotion and gender, Hess, Adams, and Kleck (2004) added male and female hairstyles to the same, androgynous faces (both real and artificial) which varied in emotional expression. They asked participants to rate these faces on their emotional intensity. Since gender stereotypes link females to happiness and males to anger, a stereotyping account of gender and emotional overlap would predict that the faces constructed to look female would look happier than their male counterparts; and vice versa for anger intensity. However, the 'female' faces actually looked more angry and less happy than the 'male' faces (see also Becker et al., 2007). This could be due to a top-down stereotypic contrast effect, as people adjust their judgement in reaction to gender expectations. Alternatively, as the authors suggest, by using androgynous faces and thus removing strong facial cues to gender along with any accompanying misleading emotional cues, this allows the true emotional facial cues to be seen more clearly (i.e. a bottom-up effect).

Clearly, these explanations are not easy (and perhaps impossible) to completely untangle, since, as Hassin and Trope (2000) and Freeman and Ambady (2011) have argued, prior expectations probably interact with facial cues in a complex recurrent neural dialogue. It is also potentially artificial to distinguish between 'conceptual' prior expectations and 'perceptual' facial cues (e.g. Barrett, Lindquist, & Gendron, 2007). Nevertheless, we can examine how perceivers' impressions of faces may differ depending on the social group membership of the face, as outlined in the **Issues investigated** section.

Studies have also started to examine how individual differences in prejudice affect facial categorisation, and associated trait judgements (Dotsch, Wigboldus, Langner, & van Knippenberg, 2008; Dotsch, Wigboldus, & van Knippenberg, 2011; Hugenberg & Bodenhausen, 2003, 2004). Although examining individual differences is beyond the scope of this review, this work again highlights close associations between top-down social information (in this case, individual differences in prejudice) with bottom-up

visual cues in the face. Likewise, there is a considerable literature which examines how categorisation affects the recognition of individual targets (e.g. see the substantial literature on the *other race effect*: Hugenberg, Young, Sacco, & Bernstein, 2011 for a review). Again, this is outwith the scope of this review but serves to highlight a recent move to integrate social psychological insights (in this case, perceiver motivation) with face perception (specifically, facial identity recognition).

Finally, other studies have also considered how faces belonging to different categories receive different resulting social impressions (e.g. through neural network modelling: Zebrowitz, Bronstad, et al., 2007; Zebrowitz et al., 2010; Zebrowitz, Kikuchi, et al., 2007). This was covered in part in the sections on facial cues (i.e. age, gender and race). Importantly, these studies tend to focus on one impression judgement at a time, as highlighted in the section on facial cues; moreover, they do not take into account the recent trustworthiness by dominance model of Oosterhof and Todorov (2008). An account that examines how Secord's (1958) categorisation mechanism fits with the dimensional model of facial first impressions is currently lacking from the field. This will be returned to in the next section.

1.5.4. Summary of previous research

In concentrating on the objective overlap between veridical and overgeneralised cues in faces, we have developed a sophisticated understanding of the features of the stimulus face that trigger social impressions. However, in doing so, the face perception literature has somewhat neglected the insights provided by social psychological investigations of the perceptions of groups (such as males and females) and associated stereotyping (an argument which is also made by Quinn and Macrae 2011). While research has started to examine the effects of stereotyping on the categorisation of faces (for instance, as Black or White: Freeman et al., 2011) or how individual differences in prejudice affect categorisation (Dotsch et al., 2008), relatively little research has examined the effect of group categories on impressions of character traits. Where studies do examine group differences in character traits, they tend to pick one or two examples, without examining a series of trait judgements systematically or in the light of recent dimensional models of facial first impressions (e.g. Oosterhof & Todorov, 2008). The rest of this review will set out a programme of research that attempts to address these gaps in the field.

1.6. Issues investigated

1.6.1. Testing dimensions of facial impressions with ambient images

As outlined in the **Ambient images** section, a recent development in face perception has been to use highly variable images of faces, in order to more naturalistically represent the variability in face photographs that we are routinely exposed to (for example, while browsing online). A promising development in facial first impressions research was to model facial impressions with two key dimensions, trustworthiness and dominance (Oosterhof & Todorov, 2008). This has already proved invaluable in narrowing down the focus of future research and providing a theoretical background to the study of facial first impressions. One of the strengths of Oosterhof and Todorov's (2008) approach is that they used carefully standardised facial stimuli (including computer-generated FaceGen images). However, this is also a limitation since it is currently unclear whether their conclusions will remain with the application of more highly variable images. This demonstration is necessary to validate this approach before such models can be widely used. **Chapter 2** of this thesis will attempt to test Oosterhof and Todorov's (2008) model with ambient images.

1.6.2. Integrating social groups with the key dimensions of facial impressions

Second, while studies are starting to build links between the face perception and social literatures, there is still much more to understand. For example, few studies have directly examined the effect of different facial groups and associated stereotypes on resulting facial impressions. Instead, most research focuses on individual differences in perceivers, either to index stereotyping (e.g. Johnson et al., 2012), or prejudice (e.g. Dotsch et al., 2008, 2011; Freeman & Ambady, 2011; Hugenberg & Bodenhausen, 2003, 2004). Moreover, the studies that do link face perception and stereotyping overwhelmingly focus on the process of social group categorisation rather than facial character impressions (see Freeman and Ambady 2011; Quinn and Macrae 2011 for reviews). Often, studies focus on one facial judgement at a time (Dzhelyova et al., 2011).

What is currently missing from the field is an approach that takes advantage of the important theoretical step forward made by Oosterhof and Todorov (2008) and examines how the key dimensions of trustworthiness and dominance differ depending on the categorical group that the face belongs to; or indeed, whether these dimensions are equally important for faces from different category groups. Moreover, as explained previously, the stereotype content model (SCM) bears a strong resemblance to the model of facial first impressions outlined by Oosterhof and Todorov (2008), but the links between these models of face and group impressions have yet to be tested. These issues will be tackled in **Chapter 3** of this thesis.

1.6.3. Investigating qualitative changes in facial impressions

There are also more subtle effects in the social psychological literature that have yet to be thoroughly investigated with regards to facial first impressions. One such is the idea that stereotyping can induce qualitative changes in meaning (Kunda et al., 1997; Thagard & Kunda, 1998). For example, Kunda, Sinclair and Griffin (1997, p720) have suggested that *"the same traits can take on different shades of meaning when applied to different individuals or groups"*. They demonstrated that participants inferred that trait aggressiveness involved verbal sparring when the category group 'lawyers' was implicated; but when 'construction workers' were being described, participants' thought about physically aggressive behaviours instead (Read & Miller, 1998; Vonk, 1994). Yet, changes in meaning due to social categorisation have apparently not been investigated for facial trait impressions.

One obvious example of a change in meaning is a change in evaluation (i.e. a positive or negative appraisal of a particular trait). Oosterhof and Todorov (2008) have argued that the first, trustworthiness dimension is strongly linked to valence, and that the second dominance dimension is orthogonal to this. However, social psychological studies have repeatedly shown that dominance is interpreted in different ways for male and female targets, with female targets being negatively evaluated for behaving dominantly (Amanatullah & Tinsley, 2013; Madeline E. Heilman & Okimoto, 2007; Rudman, 1998; Rudman & Glick, 1999, 2001; Rudman, Moss-Racusin, Glick, & Phelan, 2012). We might therefore expect that the dominance dimension is also evaluated differently for male and female faces, alongside broader qualitative changes

in meaning as the gender of the face changes. For instance, spontaneous impressions of faces may vary depending on the gender category of the face. These ideas are examined in **Chapter 4** of this thesis.

1.6.4. Investigating cultural differences in facial impressions

Finally, one key implication made by Oosterhof and Todorov's (2008) claim that the dimensions are the by-products of evolutionary mechanisms for threat detection, is that the dimensions of trustworthiness and dominance as judged in faces should then be universal across the culture of the perceiver and the race of the face. **Chapter 5** of this thesis will therefore broaden the focus of the field of first impressions to include a preliminary examination of a non-Western cultural group; specifically, by examining Chinese perceivers' facial impressions.

1.7. Summary

From the studies summarised in this literature review, it is clear that this is an exciting time to be studying social face perception. Recent models of facial first impressions have tied together disparate findings and multiple potential cues to facial first impressions, to highlight two dimensions in particular: trustworthiness and dominance (Oosterhof & Todorov, 2008). While research in face perception has traditionally set aside the effect of social group and stereotyping, more recent studies have looked at the interplay between top-down, conceptual knowledge of groups with a stimulus face (for example, in racial face categorisation: Freeman et al., 2011). Moreover, while many classic social psychological findings have rested on verbal labelling of stereotyped groups, there is a move towards applying social theories to more ecological stimuli (Gilbert & Hixon, 1991; Quadflieg & Macrae, 2011). These efforts are starting to result in sophisticated models that include both perceptual stimuli and stereotyping influences (e.g. Freeman & Ambady, 2011). Yet, many questions remain, and many social psychological insights have yet to be applied to the perception of faces.

Together, this thesis attempts to integrate insights from social psychology with face perception models of first impressions. The central argument made will be that although dimensional models have been important in elucidating key aspects of facial

impressions, it is also vital that research on facial first impressions also includes an investigation of categorical groups, such as the gender or culture of the face or perceiver.

Chapter 2. Social inferences from ambient images generate a three-dimensional model

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118. doi:10.1016/j.cognition.2012.12.001

2.1. Abstract

Three experiments are presented that investigate the two-dimensional valence/trustworthiness by dominance model of social inferences from faces (Oosterhof & Todorov, 2008). Experiment 1 used image averaging and morphing techniques to demonstrate that consistent facial cues subserve a range of social inferences, even in a highly variable sample of 1000 ambient images (images that are intended to be representative of those encountered in everyday life, see Jenkins et al., 2011). Experiment 2 then tested Oosterhof and Todorov's two-dimensional model on this extensive sample of face images. The original two dimensions were replicated and a novel 'youthful-attractiveness' factor also emerged. Experiment 3 successfully cross-validated the three-dimensional model using face averages directly constructed from the factor scores. These findings highlight the utility of the original trustworthiness and dominance dimensions, but also underscore the need to utilise varied face stimuli: with a more realistically diverse set of face images, social inferences from faces show a more elaborate underlying structure than hitherto suggested.

Keywords: "social inferences" "face perception" "first impressions"

2.2. Introduction

2.2.1. Current face evaluation models

We readily infer character traits from faces: indeed, 75% of people in one poll believed that you can gain some information about a person's character from their face (Hassin & Trope, 2000). These judgements can have important consequences: for example, the perceived competence of faces can influence election outcomes (Antonakis & Dalgas, 2009; Todorov et al., 2005). Consequently, it is important to understand why people judge faces in this way and what underlies these judgements.

Recently, a substantial step forward in the field of social facial attributions has been the introduction of a two-dimensional model to elucidate an underlying structure to face evaluations (Oosterhof & Todorov, 2008). Briefly, the authors asked participants to infer traits from faces, then applied principal components analysis (PCA), which reduced the trait judgements made into two underlying dimensions: trustworthiness/valence and dominance. Oosterhof and Todorov (2008) argue that these dimensions are fundamental in first impressions because they relate to the appraisal of threat. The trustworthiness/valence dimension concerns perceived *intention* to help or harm, and is based on an emotion generalisation; so that faces which appear angry are perceived as untrustworthy and therefore to be avoided, while faces which appear happy are viewed as trustworthy and approachable (Todorov, 2008; Zebrowitz et al., 2010). The dominance dimension, on the other hand, is based on perceived *ability* to carry out any helpful or harmful intentions. Underlying this inference are judgements of physical capability, maturity and masculinity (Fink et al., 2007; Oosterhof & Todorov, 2008).

This two-dimensional account has the potential to bring together a range of observations relating to different perceived traits, and thus offers a powerful theoretical integration (Bruce & Young, 2012). To cross-validate their model, Oosterhof and Todorov (2008) collected trait ratings on 300 computer-generated faces. The principal components for these faces' physical attributes were known, allowing Oosterhof and Todorov (2008) to map the perceived trait dimensions onto the 'face' space defined by these physical dimensions, with the average face centred at the origin (based on procedures by Blanz & Vetter, 1999). In support of the model, faces higher than average on the trustworthiness dimension appeared to smile, and those lower appeared increasingly

angrier; while increasingly dominant faces looked more mature, masculine and darker (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011). Moreover, the faces generated to fall on these two dimensions were indeed perceived by a new sample of participants to vary on trustworthiness and dominance (Oosterhof & Todorov, 2008).

There is also considerable independent evidence supporting this two-dimensional model. For example, Boothroyd, Jones, Burt and Perrett (2007), and Walker and Vetter (2009), carried out similar analyses on social judgements of faces and also broadly found two equivalent dimensions with similar underlying cues. Furthermore, two similar dimensions of 'warmth' and 'competence' have consistently been shown to be important within a wide range of social and personality research, such as in describing how people perceive cultural groups (Cuddy et al., 2008; Leary, 1957; Vigil, 2009; Wiggins, 1979; Wojciszke, 1994).

Previous authors have proposed that the trustworthiness and dominance dimensions have evolutionary significance; since being able to evaluate conspecifics in terms of their intentions (threatening or otherwise) and associated capabilities, and thus appropriately approach or avoid them, is crucial for survival (e.g. Oosterhof & Todorov, 2008; Watkins, Jones, & DeBruine, 2010). From this evolutionary standpoint, it is perhaps surprising that attractiveness does not play a greater role in existing face evaluation models, since it is clearly related to fundamental mechanisms of sexual mating and selection with a long evolutionary history (Buss & Schmitt, 1993; Little et al., 2011; Thornhill & Gangestad, 1999).

In Oosterhof and Todorov's (2008) model, attractiveness is largely dependent on the trustworthiness dimension, but includes to a lesser extent an influence of dominance. The subsidiary emphasis to attractiveness in the model is surprising not only from the evolutionary theoretical perspective but also in light of substantial evidence regarding the importance of facial attractiveness in first impressions (Dion et al., 1972; Little et al., 2011). Indeed, research into the *what-is-beautiful-is-good* effect has shown that physically attractive faces are ascribed other positive attributes such as sociability (Dion et al., 1972), suggesting that attractiveness perceptions could also be a fundamental dimension underlying social inferences from faces. Moreover, although in the two-dimensional Oosterhof and Todorov (2008) model attractiveness largely loads on the

trustworthiness factor, a meta-analysis investigating the strength of the *what-is-beautiful-is-good* effect found that attractiveness is not especially linked to trustworthiness or other morality related judgements (Eagly et al., 1991).

Likewise, theoretical models from the romantic partner preferences literature find a separate attractiveness dimension in addition to warmth-trustworthiness and status dimensions (Fletcher, Simpson, & Thomas, 2000; Fletcher, Simpson, Thomas, & Giles, 1999). In summary, there is considerable evidence for the importance of attractiveness in first impressions, although as a perception which is distinct from threat-related judgements.

2.2.2. Ambient images

Despite the successes of Oosterhof and Todorov's (2008) approach, it is important to note that in building such models, the previous studies have mostly employed tightly controlled, highly homogenous stimuli. Highly controlled stimuli, of course, offer the ability to precisely manipulate and examine facial parameters. Moreover, by minimising noise, subtle effects can be investigated. However, this approach necessarily ignores the considerable face variation that exists in the natural world. In doing so, it leaves open the possibility that other factors, such as attractiveness, might also influence the perception of more naturalistic stimuli in important ways.

Indeed, Jenkins, White, Van Montfort and Burton (2011) have recently argued for the importance of preserving this natural face variability to better understand identity recognition and within-identity variation (see also Burton et al., 2011). One way to maintain such variability is to sample publically available, pre-existing photographs from the internet. Jenkins and colleagues (2011) term these highly varying photographs *'ambient images'* to reflect the fact that they preserve something of the diverse conditions under which we naturally see faces.

Here, our goal is to further examine first impressions of faces and, specifically, to test Oosterhof and Todorov's social evaluation model with ambient images of different identities. Our ambient images are photographs of 1000 different faces collected from the internet, which have been deliberately chosen to display wide-ranging ages,

expressions, poses, and levels of health, and include facial hair and paraphernalia such as piercings or glasses (Santos & Young, 2005, 2008, 2011). Allowing these cues to vary reflects the wide range of faces we see in everyday life and thus allows a strong test of the utility of the valence/trustworthiness and dominance dimensions; as well as allowing other potentially important dimensions to emerge.

When considering such a variable face sample, another issue arises. At the present, it is not entirely clear how consistent are the cues underlying social impressions from faces. It is possible that with a more naturalistic sample, inferences from faces might be cued by multiple different facial attributes rather than by consistent cues. For example, both attractiveness (Zebrowitz & Rhodes, 2004) and the wearing of glasses (Leder, Forster, & Gerger, 2011) individually cue intelligence, but it is unclear whether their effects remain in combination, to form a kind of 'facial intelligence prototype', or rather, show more complicated interactions or even cancel each other out. If multiple cues are inconsistent, indicating possibly different routes to the same trait judgement, then attempting to model these cues as lying on a small number of unitary dimensions would seem to have less utility.

Moreover, at present the models largely only consider physical cues; yet, social or cultural stereotypes should also affect trait judgements (e.g. the wearing of glasses as indicating intelligence: Leder et al., 2011). A more naturalistic sample should preserve more of this information, allowing us to determine if the dimensions can account for these stereotypes as well as physical features.

In summary, modelling social inferences from faces has been a valuable and influential technique. However, given this, it is important that the model generalises to naturally varying faces, and that the model assumptions are stringently tested.

2.2.2.1. Research aims

The current set of studies utilised a database of ambient images consisting of 1000 photographs of Caucasian adult faces taken from public sources on the internet (Santos & Young, 2005, 2008, 2011). These photographs have been deliberately chosen to display faces with wide-ranging ages, expressions, poses and health; and they include

paraphernalia such as glasses, facial hair and piercings. Image characteristics including camera type and angle, lighting and background also vary. The database aims to provide 'snapshots' of brief encounters as variable as those we encounter in real life. This contrasts with previous modelling work, which has mostly used standardised photographs. Furthermore, it is the largest sample of natural face photographs that has been applied to the building of trait-face models so far.

Experiment 1 first assessed the consistency of the cues underlying facial trait inferences by utilising image averaging and morphing techniques. In brief, faces in the database rated as high and low on particular social traits were averaged together to create putative high or low prototype images for each trait. This technique is optimal for testing cue consistency because it ensures that only those attributes that consistently cue trait inferences (i.e. attributes that are present in the majority of the contributing face photographs) will be brought forth in the averages. If inconsistent features cue trait inferences, then the resulting averages will average over these features and will therefore fail to be perceived as predicted. To further demonstrate that the cues remaining in the high and low prototype images are valid signals of the trait in question, the high and low prototypes were morphed between, to form linear continua, which were rated on their respective traits.

In Experiment 2, the two-dimensional model (Oosterhof & Todorov, 2008) was tested on the ambient image database itself, by investigating the factor structure of traits rated directly from the 1000 images. Finally, in Experiment 3, image averaging and morphing techniques were used again, this time to cross-validate the three-dimensional factor structure that emerged from Experiment 2. This is the first time that both the texture (reflectance) and the shape of the three dimensions have been visualised directly, rather than using individual traits as proxies, since the averages were built from faces lying high or low on the dimensions themselves (see also Said, Dotsch, & Todorov, 2010 for a direct manipulation of face shape along the valence dimension).

2.3. Experiment 1

Experiment 1 investigated how consistent the facial cues underlying trait inferences are, given a starting sample of ambient face images. This has important consequences, because if inferences from faces are cued by multiple inconsistent facial attributes, then dimensional modelling is less advantageous. To investigate this, averaging and morphing techniques were employed. These are ideally suited to answering this question since, by averaging across exemplars, they preserve only the cues that are consistent across many different faces.

Although there is already a large literature that has used face averaging techniques to build averages of faces rated on a wide range of characteristics (see Little et al., 2011 and Tiddeman, Burt, & Perrett, 2001, for reviews), these studies mostly use images whose properties are tightly constrained. For example, a common technique in the attractiveness literature is to average full-face photographs of young adults with neutral facial expressions taken under standard lighting conditions. Such methods have the advantage of delivering highly controlled stimuli to test specific hypotheses, but they leave open the possibility that other cues might be available in the natural environment. Our study is the first to apply image averaging to ambient images in order to investigate a wide range of perceived social facial characteristics (including trait impressions).

In summary, Experiment 1 sought to extend previous work by estimating how consistent are the cues that underlie social inferences of intelligence, trustworthiness, dominance and confidence; given a highly variable, ambient image sample. In addition to averaging images to reveal underlying traits, we sought to provide converging evidence for face-trait cues by using a morphing procedure to show that each trait could be varied across a continuum.

In order to achieve this, face averages were created from ambient face photographs perceived as being high or low on the four inferred traits of intelligence, trustworthiness, dominance, and confidence. These were chosen to sample evenly throughout the two-dimensional space proposed by Oosterhof and Todorov (2008). The three physical attributes of age, sexual dimorphism (feminine-masculine) and attractiveness were also included to verify that these highly variable stimuli were indeed

still able to be manipulated as previous research suggests (e.g. Little et al., 2011; Tiddeman et al., 2001). The high and low averages for each attribute were then morphed between in order to create seven morphed continua, which could be examined to see if they varied systematically on their manipulated attributes.

2.3.1. Method 1

2.3.1.1. Initial ratings collection

The ambient image face database consists of photographs of 500 male and 500 female adult Caucasian faces taken from the internet (Santos & Young, 2005, 2011, 2011). The photographs in this database are standardised to be 150 pixels in height (approximately 5cm on screen) and have been cropped around the head and shoulders to minimise the background. Only non-famous Caucasian adults are represented. Non-Caucasian faces were deliberately excluded (in keeping with other models: e.g. Oosterhof & Todorov, 2008) to avoid the impact of other race effects (Hugenberg et al., 2011; Rossion & Michel, 2011), which could distort facial perceptions. All other variables in the database have been deliberately left unstandardised, to capture a naturalistic representation of the varying influences that might contribute to first impressions. These include facial characteristics such as age, expression, pose, health; facial hair, glasses and piercings; and image characteristics including lighting, background, camera type and angle.

Ratings of trustworthiness, approachability, degree of smiling, attractiveness, intelligence, dominance, sexual dimorphism, skin tone, confidence, aggressiveness, age and babyfacedness were collected and used in the following experiments. The attributes of trustworthiness, approachability, degree of smiling, dominance, skin-tone, sexual dimorphism, attractiveness, intelligence, confidence and aggressiveness were included based on their importance in previous facial modelling studies (e.g. Boothroyd et al., 2007; Oosterhof & Todorov, 2008; Walker & Vetter, 2009) and so that the hypothesised valence/trustworthiness by dominance space would be fully sampled. Perceptions of age, attractiveness, babyfacedness and health were also collected due to their substantial importance in other face perception studies (e.g. Little et al., 2011; Thornhill & Gangestad, 1999; Zebrowitz & Montepare, 1992).

In total, 50 participants (25 female and 25 male; mean age approximately 24 years) rated the ambient images. Participants provided written informed consent to procedures that were approved by the ethics committee of the University of York psychology department and were tested in a quiet room at various time points and locations on either a PC computer or laptop. Participants were told that they were taking part in a study of first impressions. A minimum of six independent participants rated each trait and all participants rated all 1000 face photographs on a given trait.

To minimise carryover (Rhodes, 2006), traits were rated in separate blocks. Carryover effects were not therefore considered a significant problem, due to the large number of faces that were rated in each block (1000). The order of the traits was counterbalanced across groups and participant sex, and within each block, the photographs were randomly presented. Before each block, participants were given a practice run of 10 faces randomly selected from the database.

On each trial, participants saw one photograph with a Likert scale (1-7) presented underneath. Two labels described the Likert scale for participants, with 1-7 anchored as: no smile-big smile; (very) pallid-tanned; young adult-old adult; feminine-masculine; unattractive-attractive; maturefaced-babyfaced; unhealthy-healthy; unintelligent-intelligent; unconfident-confident; nondominant-dominant; untrustworthy-trustworthy; unapproachable-approachable; or nonaggressive-aggressive. Participants pressed the number key that corresponded with their rating and the next face photograph appeared after a blank interval of approximately 750ms. Participants were given as much time as they wanted but were encouraged to go with their 'gut instinct' (Todorov et al., 2005).

2.3.1.2. Participants

12 participants (6 female and 6 male; mean age: 21.42 years) volunteered to take part in Experiment 1 in return for course credit. Participants provided electronic informed consent to procedures that were approved by the ethics committee of the University of York psychology department. Participants did not take part in the other currently reported experiments.

2.3.1.3. Stimuli & design

In the first step, we averaged together the 20 face photographs rated highest to create a high face average, and the 20 faces rated lowest to create a low face average, for each of the attributes of age, sexual dimorphism (feminine-masculine), attractiveness, intelligence, trustworthiness, dominance and confidence **(see figure 2.1.**). The face averages were constructed using Psychomorph software (version 4: Tiddeman et al., 2001). In brief, 179 fiducial points were marked on each face photograph to define the face shape. The software then averages the vectors formed by these points; warps (aligns) the corresponding image textures/colours to this average shape, and finally, averages the aligned textures together (see Tiddeman et al., 2001 for further details).

Figure 2.1. Morphed continua for age, sexual dimorphism, attractiveness, intelligence, confidence, trustworthiness and dominance. The left and rightmost faces are face averages constructed from averaging the 20 highest and lowest scoring faces for each trait. The faces in between are morphed between the endpoints in steps of 10%.

Some photographs were excluded from some or all average face images. Face photographs that were present in more than one high or low group were removed from the trait groups they contributed least to, and substituted with the next highest rated photograph. This was done in order to prevent a few face photographs dominating the face averages, since the face averages only rely on a comparatively small number of faces. At this point, two of the images were discovered to depict celebrities and three seemed, on closer inspection, to be non-Caucasian. These images were also substituted to avoid familiarity and the other race effect, which were not the current research aims. Finally, five faces that could not be delineated satisfactorily due to poor image quality were also substituted. The averages were standardised to be 400 pixels in height but all other variables were free to vary.

In a second step, the pair of average images rated high and low on a given trait were morphed between in steps of 10%, so that a morphed continuum consisting of 11 different averages was created for each of the seven traits (resulting in 77 face averages in total: **see figure 2.1.**). These continua allowed us to assess, using correlations, whether the averages were reliably perceived as changing on their respective manipulated traits.

2.3.1.4. Procedure

Participants were tested in a quiet room on a PC computer running E-Prime software (version 2; Psychology Software Tools, Pittsburgh, USA) and were told that they were taking part in a study of first impressions.

All participants rated all 77 face averages on all seven traits. Each trait was rated in a separate block; the order of the blocks was randomised and the order of the face averages was randomised within a block. While the use of the same participants to rate all traits means that carryover could inflate the correlations between the traits (Rhodes, 2006); the main aim of this experiment was to demonstrate the reliable facial manipulation of given traits, not to examine the correlations between them. Before each of the seven blocks, participants were given a practice run of six other average faces.

On each trial, participants saw a face average, with a Likert scale (1-7) presented underneath. Two labels described the scale, so that 1-7 always represented: (very) young adult-old adult, feminine-masculine, unattractive-attractive, unintelligent-intelligent, untrustworthy-trustworthy, nondominant-dominant and unconfident-confident. All other aspects of stimulus presentation were as the initial rating study.

2.3.2. Results 1

2.3.2.1. Reliability

In order to justify modelling at the face level (Todorov & Oosterhof, 2011), Cronbach's alpha was computed for each of the initial trait ratings of the 1000 ambient images. Importantly, the thirteen initial trait ratings on the ambient image database demonstrate good reliability with all alphas above .7 (Nunnally, 1978). Reliability was also calculated for Experiment 1, for each of the seven morphed trait continua. These average image ratings also demonstrate good reliability, with all alphas above .7.

2.3.2.2. Trait validation

In order to ascertain whether the traits could be reliably manipulated, correlations were computed between the obtained ratings for a given trait against the manipulated level of that trait (1-11 linear scale) for each of the seven morphed continua separately. In every case, the correlation was significant (all $n = 11$, $p < .001$) and high: age ($r = .97$); sexual dimorphism ($r = .99$); attractiveness ($r = .99$); intelligence ($r = .97$); confidence ($r = .96$); trustworthiness ($r = .93$) and dominance ($r = .94$).

Following Hönekopp's recommendation (2006), correlations were calculated between the predicted and obtained ratings for each of the twelve participants separately. These were significant (all $n = 11$, $p < .05$) for eight participants for the dominance continua; nine participants for trustworthiness and intelligence; eleven participants for confidence and all participants for age, sexual-dimorphism and attractiveness.

Cross-correlations across different traits and morphed continua were not examined here because the main purpose was to ascertain the reliability of single traits. Instead, the structure of the face trait space was examined in Experiments 2 and 3.

2.3.3. Discussion 1

Importantly, it is evident that the morphed continua were viewed as expected in terms of their respective manipulated traits, as evidenced by the high correlations between the manipulated attributes and the participants' ratings. Therefore, it seems that there are features in faces that do reliably cue social inferences, even given a highly variable initial sample of images. This consistency was not only true at the face level, but also held for the majority of individual raters.

It is also clear that the averages do indeed appear subjectively to change on their manipulated traits (**figure 2.1.**). Indeed, while the averages are not as controlled as those from previous research using more homogenous initial stimulus sets (e.g. Penton-Voak et al., 2006) it is striking just how clearly trait cues none the less emerge. For example, the skin-tones of the feminine and low dominance averages are lighter than their masculine and high dominance counterparts (Oosterhof & Todorov, 2008). Also interesting are cues which emerge but have not yet been integrated into the face evaluation modelling approach: for example, the high intelligence and low attractiveness face averages appear to have glasses, agreeing with previous stereotyping research (Leder et al., 2011; Thornton, 1944). The current study adds to this previous research by providing converging evidence for these cue-trait links from the averaging and morphing procedures.

At this point, it is worth noting that data-driven, 'reverse correlation' methods have recently been used to examine cues, by associating social inferences with artificial faces/feature changes (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011; Walker & Vetter, 2009) or random noise patterns superimposed on a 'base' face (Dotsch & Todorov, 2011). Indeed, the current morphing method acts like a kind of reverse correlation since the initial set is unbiased, and participants, not the researchers, drive what emerges in the face averages (cf. Todorov, Dotsch, Wigboldus, & Said, 2011). However, morphing also has the advantage of examining all features naturally and consistently present in these combinations in the population, including socially mediated features (e.g. glasses).

Finally, while the attributes were manipulated separately, there seemed to be similarities between the trait impressions. This can be seen in the face averages; for example, the trustworthiness and dominance continua appear also to change in sex. Experiment 2 and 3 systematically examined these cross correlations, with the aim of modelling the structure of these impressions.

2.4. Experiment 2

Experiment 2 aimed to test the Oosterhof and Todorov (2008) model using our large database of ambient images. Thirteen attributes were chosen (as described in the **Initial ratings collection)** so that enough variables could potentially load on each factor to make them meaningful (Kline, 1994). As explained previously (see the section on **Initial ratings collection),** the attributes of trustworthiness, approachability, degree of smiling, dominance, skin-tone, sexual dimorphism, intelligence, confidence and aggressiveness were included to index the hypothesised valence/trustworthiness by dominance space (e.g. Oosterhof & Todorov, 2008); and perceptions of age, attractiveness, health and babyfacedness were also collected due to their importance as described elsewhere in the face perception literature (e.g. Little et al., 2011; Thornhill & Gangestad, 1999; Zebrowitz & Montepare, 1992).

Since social first impressions have not been examined on a face stimuli set as varied as the current one, it puts the two-dimensional model to a strong test: despite the high variability of the images, our choice of traits should be able to pick up on the two hypothesised dimensions if they were present. This is because, according to previous work, trustworthiness, approachability, and degree of smiling should load on the valence/trustworthiness factor; and dominance, skin-tone and sexual dimorphism on the dominance factor (Boothroyd et al., 2007; Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011; Walker & Vetter, 2009).

As well as testing the two-dimensional model, our approach allows novel dimensions to emerge, and based on the various previous points regarding the importance of attractiveness in first impressions, this was an obvious candidate factor.

2.4.1. Methods 2

This experiment was carried out on the ratings of the 1000 ambient image photographs. A factor analysis was chosen to model the structure of face trait space, as a factor analysis is preferred to principal components analysis (PCA) for model building and structural investigation (Borsboom, 2006; Kline, 1994). This is because factor analysis attempts to model the structure between the variables and includes an estimation of error, unlike PCA (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Rather than forcing the dimensions to be orthogonal, an oblique rotation was employed. This allowed us to assess the correlations between the dimensions.

The main analysis was run on trait perceptions at the level of the faces (that is, averaging across participants' ratings for each face photograph). The thirteen ratings entering the analysis consisted of trustworthiness, approachability, degree of smiling, attractiveness, intelligence, dominance, sexual dimorphism, skin tone, confidence, aggressiveness, health, age and babyfacedness.

2.4.2. Results 2

2.4.2.1. Reliability

As described before, the ratings of the ambient image database demonstrate good reliability, with alphas above .7 (Nunnally, 1978). Bartlett's test of sphericity indicated that the correlations were large enough that a factor analysis was appropriate; X^2 (105) = 10,777, $p < .001$ (**see table S2.1. in the supplementary material**, for the correlational matrix).

2.4.2.2. A three-dimensional model

First, a principal axis factor analysis was carried out without rotation in order to determine the number of factors to be extracted (Fabrigar et al., 1999). Four criteria were utilised to determine this, in an attempt to be as objective as possible. These criteria included the traditional Kaiser's criterion and scree test (Kline, 1994). However, since these criteria have been criticised for being arbitrary and subjective (Fabrigar et al., 1999; O'Connor, 2000), a parallel analysis (Horn, 1965) and minimum average partial analysis (Velicer, 1976) were also carried out (see O'Connor, 2000 for more

details). The first three analyses indicated that three factors were present and the minimum average partial analysis indicated that four were present. Thus, three factors were retained.

Second, the principal axis factor analysis solution was rotated, to determine the factor structure and loadings (Kline, 1994). A direct oblimin rotation was chosen to allow the factors to remain oblique. Following Kline (1994), the structure matrix was interpreted, ignoring loadings below .3 **(table 2.1.**; for further information, **see table S2.2. supplementary material).**

Table 2.1. Principal axis factor analysis: Structure matrix. These can be interpreted as correlations between the factors and variables.

Trait	**Factor 1**	**Factor 2**	**Factor 3**
Aggression	**-.94**	.21	.06
Approachability	**.91**	-.23	.21
Trustworthiness	**.89**	**-.37**	.08
Smile	**.86**	-.20	.08
Confidence	**.58**	**-.41**	**.49**
Health	**.33**	**-.87**	**.39**
Attractiveness	**.41**	**-.87**	.27
Age	.04	**.71**	**.32**
Babyfacedness	.18	**-.49**	-.14
Dominance	**-.37**	**.44**	**.82**
Sexual Dimorphism	**-.32**	**.45**	**.56**
Intelligence	**.37**	-.16	**.45**
Skin	.15	-.12	**.39**

Note: Substantial loadings (above .3: Kline, 1994) are highlighted in bold.

The first factor appears to replicate the valence/trustworthiness factor (Oosterhof & Todorov, 2008) with high loadings from approachability, trustworthiness and degree of smiling. There is also a high negative loading from aggressiveness. An appropriate factor name might thus be 'approachability'. The third factor also appears to replicate the previous dominance factor (Oosterhof & Todorov, 2008), with dominance, sexual dimorphism (increasing masculinity) and age contributing as expected. Confidence and intelligence also load highly on this factor.

However, the second factor is novel: it has a high positive loading from age, and high negative loadings from attractiveness, health, and babyfacedness. Consequently, it appears to be a negative 'age' factor, with increasing age perhaps corresponding with decreasing sexual fitness. For ease of interpretation, this factor is henceforth described in inverse form, as 'youthful-attractiveness'.

Unfortunately, after oblique rotation, one cannot determine the proportion of variance explained by the (rotated) factors. As an indication, before rotation, these three factors explained 72.38% of the variance with factor 1 contributing 37.76%; factor 2, 18.45%; and factor 3, 16.18%. Moreover, a separate orthogonally rotated PCA solution generated a similar result, with each principal component explaining 31.39%, 22.53% and 18.46% respectively. While these data cannot be directly applied to the rotated solution, this does demonstrate broad comparability with previous research.

The factor correlations are: -.33 between factors 1 and 2; .11 between factors 1 and 3; and .02 between factors 2 and 3. Thus, it appears that the approachability and youthful-attractiveness factors cluster slightly closer together than with the third dominance factor, which is almost entirely independent.

2.4.2.3. Model robustness

To ascertain the model robustness, different analyses were implemented and various traits excluded (**see table S2.3. supplementary material**). All analyses employed produced a nearly identical three-factor solution, including a PCA with orthogonal rotation, demonstrating that the current result is not dependent on the analysis, but reveals a structure present within the data. However, inferences (e.g. regarding the factor loadings) are preferable from the factor analysis (Fabrigar et al., 1999).

Interestingly, when a PCA with orthogonal varimax rotation was carried out with the solution restricted to find two factors, an approachability by dominance solution emerged. In other words, although a three-factor solution is more justified on the basis of the majority of the initial criteria, evidence for the two predicted dimensions emerged when thus constrained.

2.4.2.4. Goodness-of-fit

Confirmatory factor analyses were then undertaken to ascertain the relative goodness-of-fit of the two models (AMOS version 18; IBM software). To make testing as fair as possible, given that the original two-factor model arose outside the present study, the dataset was randomly split (each $n = 500$; balanced for face sex). The three-factor model was then replicated in one half of the data **(see table S2.3.)** and all confirmatory analyses were carried out on the other half. Multivariate normality was acceptable (Byrne, 2009) and the maximum likelihood method was employed (T. A. Brown, 2006).

The first model tested had orthogonal approachability and dominance dimensions, as the following loadings constrained to zero: smiling, trustworthiness, approachability and health on the dominance factor; and dominance, skin tone, age, and sexual dimorphism on the approachability factor (following Boothroyd et al., 2007; Oosterhof & Todorov, 2008). The first factor was scaled to trait approachability and the second to trait dominance. The second model was oblique and also included the youthful-attractiveness factor, which was scaled to trait attractiveness. For this three-dimensional model, factor loadings under .3 (taken from the first split-half) were constrained to zero.

Multiple different indices of fit were used to assess the relative fit between the two different models; including chi², the Root Mean Square Error of Approximation (RMSEA), a comparative fit index (CFI), and a predictive fit index (Akaike's information criterion, AIC) following Harrington (2008). A 3D model presented a better fit on all four indices **(table 2.2.)**. Note that while three factors always explain more of the variance than two factors within a factor analysis; within the confirmatory testing, the RMSEA and the AIC criteria take parsimony into account and all being equal, favour simpler models with greater degrees of freedom (see T. A. Brown, 2006 for a detailed description of their computation). In short, comparing models with differing numbers of factors is fair provided that one has a theoretical reason for each model (T. A. Brown, 2006). Orthogonal and oblique models were also then compared: orthogonal factors had a slightly worse fit on three out of four indices for the 2D model and on all indices for the 3D model.

Table 2.2. Goodness-of-fit for competing 2D and 3D models.

Model	X^2 (df)	associated *p*-value	RMSEA	associated *p*-value	CFI	AIC
2D orthogonal	2325 (61)	$p < .05$	.27	$p < .05$	.59	2385
2D oblique	2280 (60)	$p < .05$	.27	$p < .05$	.60	2342
3D orthogonal	1103 (55)	$p < .05$	.20	$p < .05$	.81	1175
3D oblique	991 (52)	$p < .05$	.19	$p < .05$	.83	1069

2.4.3. Discussion 2

The approachability and dominance factors found through our analyses replicate the valence/trustworthiness and dominance dimensions from previous work (Todorov, 2008). However, a novel dimension also emerged, best described as 'youthful-attractiveness'. Moreover, this three-dimensional model clearly showed a better fit than the original two-dimensional model. The finding of an additional attractiveness factor is not entirely surprising, as previous studies did not utilise such varied stimuli: without this variance, this factor was perhaps not free to emerge.

It is interesting that while in previous studies, attractiveness mainly contributed to the approachability dimension (following Todorov & Oosterhof, 2011; Walker & Vetter, 2009), here, it was powerful enough to emerge as a dimension in its own right. Seemingly, when less constrained, the visual cues that make a face appear young and beautiful are substantially different from those that make it approachable or trustworthy (or indeed, dominant). In Experiment 3, this was explored further with face averages.

2.5. Experiment 3

In the third experiment, face averages were generated to cross-validate the model in a new sample. As described previously, the averaging technique (Tiddeman et al., 2001) allows one to visualise only the properties common to the majority of faces, in this case, those lying high or low on a factor.

This visualisation of the three factors was achieved by first calculating factor scores for each face photograph, taken from the three-dimensional model emerging in Experiment

2. Then, face averages were created from the 20 highest or lowest factor-scoring face photographs for each of the three factors of interest. If the dimensions do not easily approximate individual traits but instead are cued by many and inconsistent attributes, then these will be averaged away. Consequently, participants will not perceive the face averages as predicted and the model will fail to be cross-validated. As mentioned previously, this is the first time that both the texture and shape of the facial dimensions has been visualised directly, rather than via trait proxies (see also Said et al., 2010 for face shapes directly generated from the trustworthiness/valence factor).

2.5.1. Methods 3

2.5.1.1. Participants

30 participants (15 female and 15 male; mean age: 20.43 years) volunteered to take part in return for course credit. Participants were separated into three groups, which were balanced for gender. They provided electronic informed consent to procedures that were approved by the ethics committee of the University of York psychology department and they had not taken part in any of the other currently reported experiments.

2.5.1.2. Stimuli

Firstly, factor scores representing each dimension of the 3D oblique model were computed for all 1000 original photographs using the regression method. Then, averages were formed by averaging the 20 highest and lowest factor-scoring face photographs for each dimension (**figure 2.2A.**), using Psychomorph software (version 4: Tiddeman et al., 2001). These 6 averages reflect 'prototypical' factor extremes for the three dimensions of interest. As in Experiment 1, face photographs were prevented from being in more than one average, and the averages were standardised to 400 pixels in height. All other variables were free to differ between the averages.

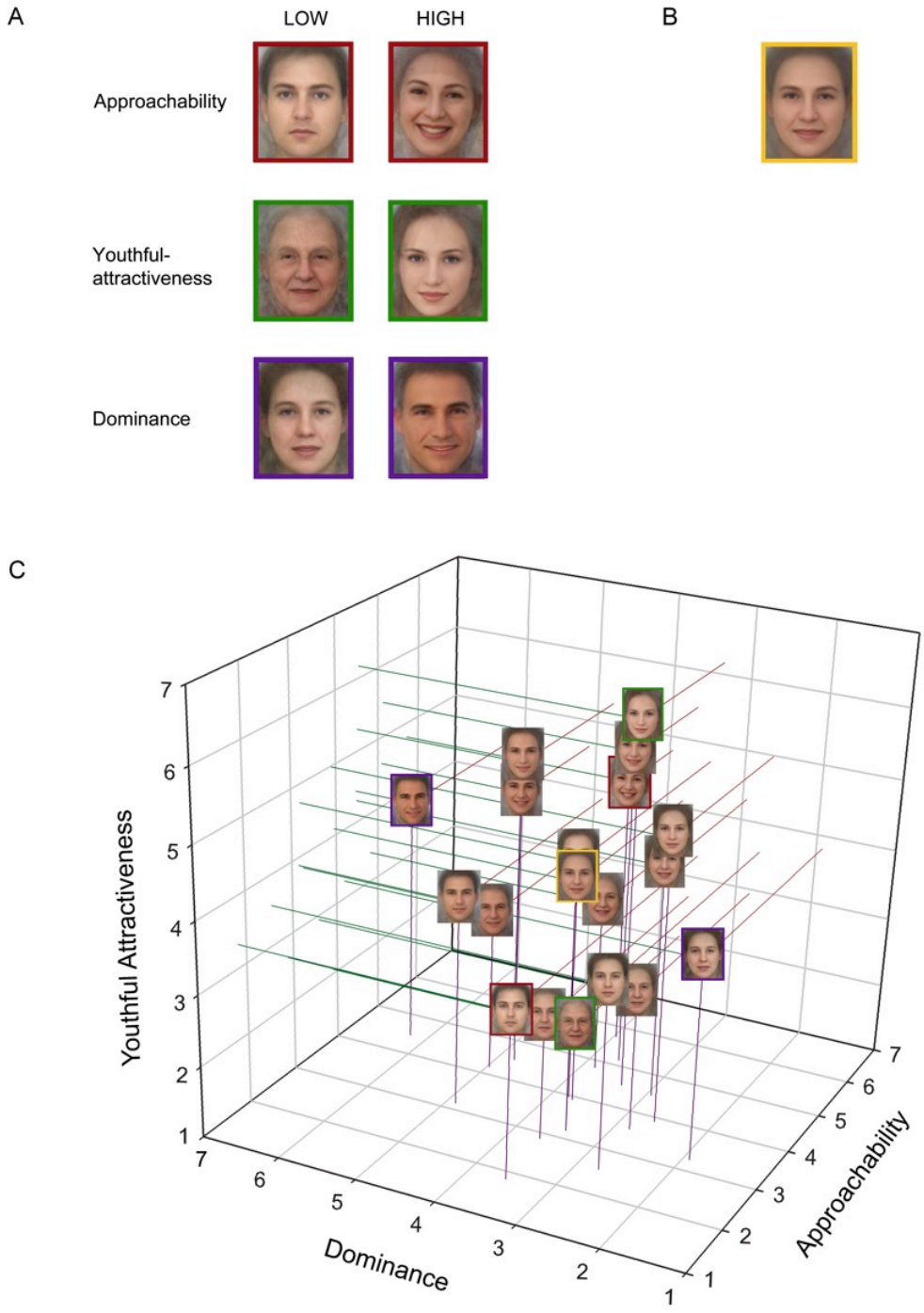

Figure 2.2. Experiment 3 stimuli: A) High and low face averages for each of the three dimensions. Each is an average of the 20 highest or lowest factor-scoring face photographs on a dimension. B) The grand-origin face average, which is an average composed of the original 6 face averages. C) 19 out of the 25 face averages created to map the three dimensions, depicted lying in the model space. Six of the face averages are not depicted because they highly overlap, being constructed to lie very close to the origin (the three 2D origin estimates, and the three high-low pairs on a single dimension). In all parts of the figure, the high and low approachability face averages are framed in red; the high and low youthful-attractiveness in green; the high and low dominance in purple and the grand-origin in yellow.

To map the model space more fully, new averages were then generated by averaging together a) all possible pairs from the six original averages; resulting in fifteen averages which fell halfway between the model extremes; b) two high and low matched pairs from the six original averages; resulting in three two-dimensional origin averages and c) all six original averages to create a grand-origin average (**figure 2.2B.**). This resulted in 25 averages in total, which together systematically mapped the three-dimensional model space. The final set of face averages are presented in **figure 2.2C.**

2.5.1.3. Design

In order to test the model, the predicted ratings for these 25 averaged images (based on the mean factor scores of their constituent photographs) were compared with the actual ratings obtained for each averaged image. Predicted ratings were derived as follows: for the six original averages, the predicted factor score was the mean of the factor scores from the 20 individual face photographs that created that average. The other predicted factor scores were then computed by averaging relevant combinations of the six original predicted scores. Note, the factor score predictions originally took the form of a ±3 point scale centred on 0; but, for ease of comprehension, were shifted to a 1-7 rating scale by adding 4.

Regarding the obtained ratings, participants necessarily could not rate the average faces directly on the dimensions because each dimension is too complex to be rated directly, being constructed from multiple traits. Therefore, the three highest loading traits on each dimension were selected as a proxy for that dimension and participants rated the face averages on these traits instead. Specifically, a group of participants rated the 25 face averages on (inverse) aggressiveness, approachability and trustworthiness, to approximate the approachability dimension. A second group of participants rated the 25 images on health, attractiveness and (inverse) age, to approximate the youthful-attractiveness dimension. Finally, a third group of participants rated the 25 images on dominance, sexual dimorphism and confidence, to approximate the dominance dimension.

Each participant's three trait ratings were then averaged together to represent the given dimension. In order that the traits approximated the dimensions as closely as possible,

this average was weighted by the traits' factor loadings from Experiment 2 (taken from **table 2.1.)** For example, to approximate the approachability dimension, each participant's approachability, trustworthiness and (inverse) aggressiveness ratings were averaged together for each of the 25 images, weighted by these three traits' loadings on the approachability dimension. This produced an obtained 'approachability' score for each image, per participant in that group. An analogous procedure was carried out for the other two dimensions.

2.5.1.4. Procedure

Participants were tested in a quiet room on a PC computer running E-Prime software (version 2; Psychology Software Tools, Pittsburgh, USA). Participants were told that they were taking part in a study of first impressions.

All participants rated all 25 face averages. Since one of the objectives was to cross-validate the factor correlations, carryover effects (Rhodes, 2006) could be a significant problem. However, as described before, three groups of participants were used (balanced for gender), with each group rating only the three traits chosen to approximate one dimension (e.g. approachability, trustworthiness and aggression). This between-subjects design at the factor level ensured that carryover effects could not contaminate the factor correlations. Moreover, the carryover effects would not be a problem for individual participants either because each participant's three trait ratings were averaged together to give a single dimensional score per participant (as described previously). Therefore, since the participants all saw the same 25 face averages, any cross-dimensional correlations that emerge must be due to the cues within these images.

The traits were separated into three blocks per group, presented in random order, and within each block, the face averages were randomly presented. Before each of the blocks, participants were given a practice run of 10 other face averages. All other aspects of stimulus presentation were as the previous experiments.

2.5.2. Results 3

2.5.2.1. Reliability

The current trait ratings demonstrate good reliability, with alphas above .7 (Nunnally, 1978). Therefore, each participant's three trait ratings for each stimulus were combined in a weighted average to approximate the dimensions. The mean rating for each face stimulus on each dimension was then calculated by averaging the weighted averages across participants.

2.5.2.2. Model validation

To elucidate formally whether the model was cross-validated, correlations between the predicted and average obtained ratings *(N* = 25) for each dimension were computed (**figure 2.3.**). The dimensions behaved as expected, as the obtained face ratings correlated with the ratings predicted for: approachability ($r = .91$, $p < .001$), youthful-attractiveness ($r = .89$, $p < .001$) and dominance ($r = .78$; $p < .001$).

Regarding the predicted factor relationships, there was a significant correlation ($r = .50$, $p = .012$) between predicted approachability and predicted youthful-attractiveness. That is, the averages had the potential to be correlated in Experiment 3 (reflecting the original Experiment 2 factor analysis). However, no other cross-dimensional correlations were significant. Thus, while the dimensions were allowed to be oblique and the stimuli were potentially correlated, this was not strong enough to emerge as a significant pattern in the participants' actual ratings.

Finally, each of the thirty participants showed a significant correlation between their individual weighted average ratings and the predicted scores for that manipulated dimension (all $n = 25$; all $p < .05$). The specificity of the dimensions was also assessed at the individual level by calculating correlations between each dimension's predicted scores with the obtained ratings from each individual participant separately. These obtained-predicted correlations were then transformed using Fisher's *r-z* transformation and compared using independent t-tests. For all three dimensions, the individual participant correlations between a dimension's predicted ratings with the obtained ratings on that dimension, were significantly higher on average than the correlations

between that predicted dimension with either of the other two obtained ratings (all n = 20, $p < .001$).

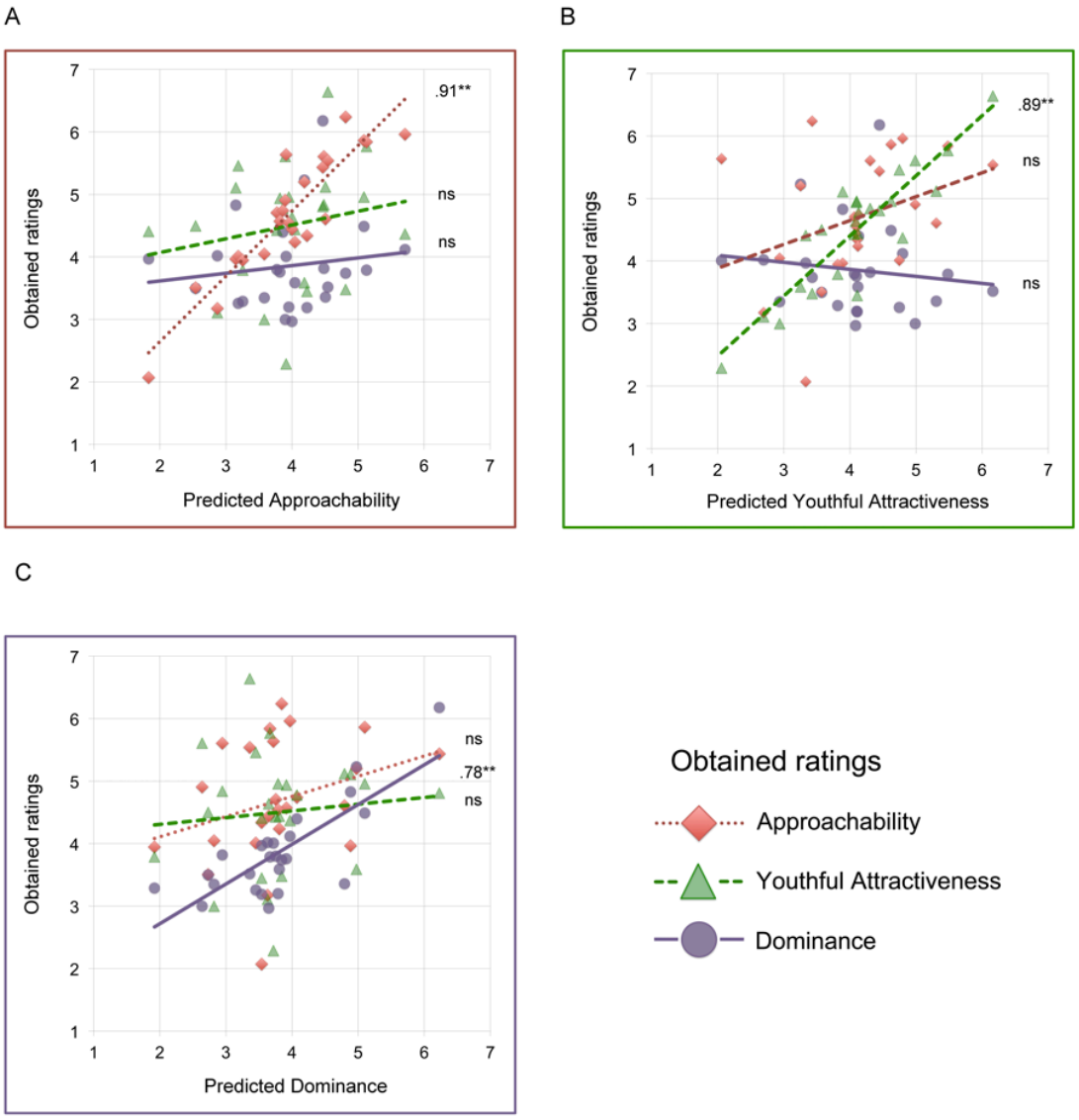

Figure 2.3. Correlations between the three sets of (averaged) obtained ratings separately with predicted approachability (A), youthful-attractiveness (B) and dominance (C). Each point represents a face average. ** $p < .001$.

2.5.3. Discussion 3

The face averages can be seen to differ on the cues which correspond to the dimension of interest, thereby acting as a qualitative cross-validation of the current three-dimensional model (**figure 2.2A.**). The high youthful-attractiveness average can be seen to be younger, healthier and more attractive than the low counterpart, although one

consequence of the averaging procedure is to smooth out skin detailing such as wrinkles, which reduces perceived age (Tiddeman et al., 2001). The high approachability average seems to be female and smiling, whereas the low approachability average appears to be male, and either neutral or negatively valenced. This corresponds with the current analyses and with previous research (Hess et al., 2004; Said et al., 2009). The high dominance average clearly looks older, less babyfaced and more unambiguously masculine than the low dominance average.

Interestingly, it was observed that approximately half of the individual face photographs that went into the high dominance average appeared young and physically fit, whereas the rest seemed older and perhaps more socially dominant. This may reflect the subtle distinction between physical and social dominance (Watkins et al., 2010). However, in general, while the individual faces going into the averages varied on many attributes (e.g. hairstyle), they also demonstrated considerable group consistency (for example, all faces which entered the high approachability group were smiling). Certainly, the cues underlying trait evaluations seem to be consistently present in faces scoring high or low on that dimension.

The dimensions also acted quantitatively as expected, with only minor exceptions (**figure 2.3.**). Importantly, the predicted scores on a given dimension were significantly correlated with the obtained scores for that dimension. This was also true at the level of the individual rater. This indicates that the dimensions can be replicated and controlled in a new sample, and that they are indeed based on consistent trait cues.

Although the face averages were not controlled to be orthogonal this was the obtained result, supporting the assumption of orthogonal dimensions. In the current experiment, the participants did not rate the face averages on more than one dimension, eliminating carryover effects. This perhaps explains the different result from the confirmatory analysis in Experiment 2, although the difference between orthogonal and oblique models was never large.

Finally, there was a slightly lower correlation between predicted and obtained dominance than the equivalent approachability and youthful-attractiveness correlations. This may be because the dominance face averages did not vary in sex as much as they

perhaps might have; this could have occurred because the factor scores (which determined the individual faces entering the face averages) are themselves only an estimate of the dimensions, or through a loss of information from the averaging procedure. Indeed, given these points, the clear cross-validation is impressive.

2.6. General discussion

In Experiment 1, averaging and morphing techniques were used to show that consistent cues subserve trait inferences made from faces, even from a starting set of 1000 highly varying, ambient image stimuli. Three dimensions of approachability, dominance and youthful-attractiveness were found to underlie trait inferences made from these faces (Experiment 2) and this three-dimensional model was then cross-validated using morphed stimuli (Experiment 3).

The most striking current result was the third novel facial 'youthful-attractiveness' factor that consistently emerged with this large and relatively unconstrained set of face stimuli. A likely reason why this factor was found here and not in previous studies lies in the wider range of ages of the faces we used as stimuli; which could support variation on both perceived age and perceived attractiveness (e.g. Thornhill & Gangestad, 1999 show that age and attractiveness are clearly linked). Other cues that may explain the emergence of this factor include textural cues that support attractiveness (e.g. Fink et al., 2001) and are likely to vary more in ambient images than in computer generated faces. The demonstration of this third factor clearly has implications for understanding human perception of faces as well as for fields beyond the academic study of human perceptions (for example, in computer graphic modelling: Arya, Jefferies, Enns, & DiPaola, 2006). When faced with a realistically diverse set of face stimuli, social inferences from faces show a more elaborate underlying structure than hitherto suggested by social face perception models.

Of course, real-world impact presents a strong argument for the importance of the youthful-attractiveness factor. For example, the cosmetic surgery industry attests to the real-life importance of these perceptions: in 2010 alone, over five million wrinkle-

reduction cosmetic treatments were carried out, contributing to an economic sector worth $10.1 billion that year (American Society of Plastic Surgeons, 2010).

There is also clear experimental evidence for increasing age being associated with decreasing attractiveness and health (Ebner, 2008; Thornhill & Gangestad, 1999). This is often explained within an evolutionary framework, in which sexual selection has equipped us with mate preferences that are highly sensitive to fitness cues such as health and age (e.g. Buss & Schmitt, 1993; Little et al., 2011; Thornhill & Gangestad, 1999). Indeed, given this substantial body of evidence, it would be surprising if sexual selection motivations did not contribute to first impressions of faces.

However, the youthful-attractiveness factor is also compatible with age stereotypes (Cuddy et al., 2008; Krings, Sczesny, & Kluge, 2011). Our participants were encouraged to use a single standard, and some of the judgements were relatively objective (e.g. age), which might increase the likelihood of stereotyping (Biernat & Manis, 1994). Indeed, this is also true for the other dimensions, as two dimensions of 'warmth' and 'competence' have also previously been shown to be fundamental in non-face areas such as the evaluation of social groups, demonstrating that the faces themselves are probably not solely driving effects (Cuddy et al., 2008; Osgood, Suci, & Tannenbaum, 1957; Vigil, 2009; Wiggins, 1979; Wojciszke, 1994). In the current set of studies, as traits or dimensions were modelled, sex clearly also changed, perhaps partially reflecting a gender stereotype. While the aim of the current set of experiments was to first of all clearly establish that these inferences exist with diverse facial stimuli, an interesting next step would be to examine to what extent these facial inferences are mediated by such stereotyping.

Although the non-face models of social perception are mostly two-dimensional, in contrast to the current results, an attractiveness factor is clearly less likely to emerge with these more abstract concepts. Nevertheless, a highly similar three-dimensional warmth-trustworthiness, status and attractiveness-vitality model emerges in the literature examining partner preferences (Fletcher et al., 2000, 1999) and the influential semantic differential model for representing attitudes (Osgood et al., 1957) also found that three dimensions were needed. The third dimension represented ‘activity’, which

bears some relation to our youthful-attractiveness factor, in as much as increasing age accompanied by decreasing health also implies decreasing activity.

Finally, with regards to this factor, it should be noted that the current raters were all relatively young. Although the current sample is comparable to previous other modelling studies (e.g. Oosterhof & Todorov, 2008), examining other age or cultural groups using this paradigm may be a point of interest for future work. Similarly, the current facial database only included faces that were considered to look Caucasian, as a first step and in keeping with previous studies (e.g. Boothroyd et al., 2007). Future research could seek to model the dimensions underlying social inferences from faces of other ethnicities.

Another important direction for future research lies in untangling the contribution of image variability (within-person) relative to facial variability (between-person variability: Jenkins et al., 2011). For example, the work of Jenkins and colleagues (2011) has shown that perceived attractiveness can vary substantially across different photographs of the same individual, raising the question of how this image variability contributes to the three dimensions we found. For face photographs, sources of image variability can be due to differences in lighting and camera properties, malleable facial characteristics such as expressions or structural differences between different faces themselves. The latter two sources of variability correspond to what Haxby, Hoffman and Gobbini (2000) termed changeable or invariant properties of faces. The approach taken here might therefore be extended in future work to ask whether different images of the same person would vary or cluster across the observed dimensions. We might expect images of the same person to vary more along dimensions that rely to a relatively greater extent on more changeable aspects of a face, such as facial expression, which is known to contribute substantially along the trustworthiness dimension (Oosterhof & Todorov, 2008). Dimensions linked more closely to relatively invariant structural characteristics of the face might in comparison show less inter-image variability. Similar effects may be expected to occur for extra-facial image properties. One strength of the ambient image approach is that it potentially renders these different sources of variability open to systematic investigation.

Despite uncovering an additional factor, it is important to emphasise that the current research none the less also found considerable support for the original two dimensions found by Oosterhof and Todorov (2008), among others. This was the case despite using different analyses, employing averaging and morphing techniques, and while utilising a larger and highly varying original face sample. This underlines the importance and wide applicability of these dimensions in the social evaluation of faces. Moreover, the assumption of dimensional orthogonality has also been shown to be highly tenable. While oblique models fitted marginally better than orthogonal ones in Experiment 2, this difference was only slight and failed to replicate in Experiment 3. This separation between the dimensions seems to go against recent claims that trustworthiness and attractiveness cues are almost identical (Xu et al., 2012). Rather, our findings are more consistent with previous modelling (Walker & Vetter, 2009) and support studies showing that facial trustworthiness and attraction can dissociate (DeBruine, 2005; Eagly et al., 1991; Etcoff, Stock, Haley, Vickery, & House, 2011). Thus, while attractiveness judgements are important in both evolutionary and current terms, they are clearly distinct from the threat-related dimensions of trustworthiness and dominance found by Oosterhof and Todorov (2008), and also in the current paper.

2.7. Conclusions

In sum, the current study developed and validated a three-dimensional model as well as demonstrating that consistent cues subserve both traits and factors. The approachability and dominance factors strongly support previous research (Oosterhof & Todorov, 2008) and the novel youthful-attractiveness factor can be interpreted either with reference to age stereotyping or in light of evolutionary psychology. As well as having theoretical and practical implications for facial trait modelling, these results further highlight the prominence of youthfulness and attractiveness perceptions in face evaluation, an issue clearly also important in the real world.

2.8. Supplementary material

Table S2.1. Intercorrelations between ratings of facial attributes.

	1	2	3	4	5	6	7	8	9	10	11
1. Age	-										
2. Aggression	-.05	-									
3. Approachability	.05	-.84**	-								
4. Attract.	-.51**	-.25**	.33**	-							
5. Babyfaced	-.51**	-.18**	.20**	.36**	-						
6. Confidence	-.03	-.45**	.53**	.58**	.10**	-					
7. Dominance	.41**	.44**	-.22**	-.18**	-.31**	.08*	-				
8. Sexual Dimorphism	.35**	.37**	-.13**	-.31**	-.22**	-.04	.82**	-			
9. Health	-.51**	-.18**	.28**	.85**	.36**	.59**	-.06*	-.17**	-		
10. Intelligence	.21**	-.24**	.37**	.37**	-.01	.39**	.17**	.04	.33**	-	
11. Trustworthiness	.03	-.83**	.78**	.46**	.17**	.51**	-.40**	-.42**	.37**	.42**	-
12. Smile	-.01	-.84**	.88**	.24**	.17**	.53**	-.28**	-.18**	.21**	.20**	.71**

Note: $n = 1000$; * $p < 0.05$, ** $p < 0.01$ (both 2-tailed).

Table S2.2. Principal axis factor analysis pattern matrix, structure matrix and commonalities.

	Pattern matrix			Structure matrix			Communalities (after extraction)
Trait	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3	
Aggression	-.99	-.10	.18	-.94	.21	.06	.92
Approachability	.92	.06	.10	.91	-.23	.21	.85
Trustworthiness	.86	-.10	-.02	.89	-.37	.08	.80
Smile	.88	.07	-.02	.86	-.20	.08	.74
Confidence	.44	-.29	.45	.58	-.41	.49	.59
Health	.01	-.88	.41	.33	-.87	.39	.93
Attractiveness	.11	-.84	.28	.41	-.87	.27	.85
Age	.25	.78	.27	.04	.71	.32	.65
Babyfaced	.05	-.47	-.13	.18	-.49	-.14	.26
Dominance	-.37	.30	.86	-.37	.44	.82	.97
Sexual Dimorphism	-.28	.35	.59	-.32	.45	.56	.58
Intelligence	.30	-.07	.42	.37	-.16	.45	.31
Skin	.07	-.11	.38	.15	-.12	.39	.17

Table S2.3. Replication of the factor structure. The three highest loadings are shown.

Analysis	Rotation	Traits included	Factor no.	Factor 1	Factor 2	Factor 3
PCA	Varimax	As main	3	- aggression approach smile	- age attractiveness health	dominance confidence intelligence
Maximum likelihood	Promax	As main	3	- aggression approachability smile	health attractiveness confidence	dominance sexual dimorph. age
Alpha factoring	Oblimin	As main	3	- aggression trust approach	dominance sexual dimorph. intelligence	health attractiveness - age
PAF	Oblimin	Age removed	3	approach - aggression smile	dominance sexual dimorph. - trust	attractiveness confidence intelligence
PAF	Oblimin	Approachability removed	3	- aggression trust smile	dominance confidence sexual dimorph.	health attractiveness - age
PAF	Oblimin	Split half 1	3	- aggression approach trust	dominance sexual dimorph. age	health attractiveness - age

Chapter 3. Convergent and divergent impressions of male and female faces.

3.1. Abstract

Prominent recent models of first impressions from faces have particularly emphasised two underlying dimensions of trustworthiness and dominance. However, it is not known whether the same dimensions would appear or are equally important for male and female faces when examined separately. Moreover, these dimensions have frequently been compared to the warmth (cf. trustworthiness) and competence (cf. dominance) dimensions in social psychological models of inter-group perception, and it has been suggested that these reflect universal dimensions of social cognition. We therefore built structural models for first impressions of male and female faces separately to investigate whether the same dimensions are universal across facial gender, and to investigate their similarity to facial and social model dimensions. Using different methods across four studies we consistently found that while perceptions of facial trustworthiness and warmth were closely related to each other, the perception of facial dominance and competence was less strongly related, especially for female faces. Taken together, our results for facial first impressions demonstrate strong similarity on the first dimension across facial and social models, with less similarity on the second dimension. Competence rather than dominance may better represent how female faces are judged.

Keywords: “first impressions “face perception” “gender” “social cognition”

3.2. Introduction

3.2.1. Dimensional accounts of facial first impressions

A considerable social cognitive literature along with anecdotal evidence throughout history attests to the power and importance of first impressions (Hassin & Trope, 2000; Olivola et al., 2014). Indeed, from just a single photograph of the face of a stranger, perceivers can rapidly and spontaneously form impressions of the character of the person depicted (Oosterhof & Todorov, 2008; Sutherland et al., in press; Willis & Todorov, 2006). Recently, dimensional approaches have become very influential in understanding these facial first impressions (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009).

For example, in one prominent study, Oosterhof and Todorov (2008) used principal components analysis to reduce a variety of spontaneous trait impressions into two underlying dimensions of trustworthiness and dominance. The first dimension closely approximates trustworthiness trait judgements, and concerns the perceived target's *intentions*, good or bad. This dimension is strongly (but not exclusively) influenced by cues to emotion, so that faces that structurally resemble angry expressions are perceived as untrustworthy and therefore should be avoided, while faces that structurally resemble happy expressions are perceived as trustworthy and therefore can be approached (Sutherland et al., 2013; Todorov, Baron, & Oosterhof, 2008; Vernon et al., 2014). The second dimension approximates dominance judgements and concerns perceived target *capability* to carry out their intentions (good or bad). Prominent cues to this dimension include facial maturity and masculinity (Oosterhof & Todorov, 2008; Sutherland et al., 2013). Oosterhof and Todorov (2008) suggest that trustworthiness and dominance are fundamental facial judgements because together they function to assess the target's threat. More recently, we have found that three dimensions subserve first impressions made to a more varied sample of ambient image faces (Sutherland et al., 2013; Vernon et al., 2014). Crucially, we replicated Oosterhof and Todorov's (2008) findings of trustworthiness and dominance dimensions, with a third youthful-attractiveness factor perhaps representing cues linked to sexual selection.

3.2.2. Facial gender

These dimensional models have provided a framework for understanding facial first impressions, have helped guide research in this field, and have opened up new avenues of exploration. However, as yet, little is known about how universal these dimensions are across different social groups. For example, it is not known whether these dimensions apply equally to male and female faces, or indeed how facial first impressions are organised within gender groups. Current studies use either both male and female faces in the same models, or only include male faces (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009) but as yet studies have not compared male and female facial first impression dimensions. Establishing whether these models apply to both male and female faces would increase their theoretical breadth, while finding facial gender differences would open up a discussion on how the social category of the face (in this case gender) might influence facial person perception.

Facial first impressions models link the facial impression dimensions to evolutionary by-products of adaptive mechanisms for judging threat and sexual selection (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Todorov, Said, et al., 2008), which might indicate that similar facial first impressions underlie the judgements of male and female faces. However, there are also reasons to believe that the faces of men and women might elicit different impressions from perceivers.

First, a substantial body of social psychological research shows that women and men are stereotypically perceived to have different character traits. Women are traditionally viewed as warm, loyal, friendly, and higher on interpersonal traits than men, while men are traditionally viewed as competitive, dominant, independent and higher on agentic traits than women (e.g. Bem, 1974; Eagly & Mladinic, 1994; Prentice & Carranza, 2002). Moreover, women and men are evaluated differently if they hold gender counter-stereotypical traits. For example, men encounter intensified proscriptive norms against being sensitive, emotional or naïve (stereotypically female traits), and women encounter intensified proscriptive norms against being controlling, arrogant or stubborn (stereotypically male traits: Prentice & Carranza, 2002). Similar gender differences are also found for facial trait evaluation: dominance is more negatively evaluated for female

faces than male faces when participants are asked to freely describe or explicitly evaluate faces (Sutherland et al., in press). Finally, women are stereotypically (and actually) linked to professions requiring social and caring abilities, such as nursing or teaching, and men are stereotypically (and actually) linked to professions requiring agentic and technical skills, such as banking or engineering (Eagly & Steffen, 1984; M. J. White, Kruczek, Brown, & White, 1989; M. J. White & White, 2006). These gender differences in socially-oriented versus agentic occupations also emerge when faces are rated on their occupational suitability (Oldmeadow, Sutherland, & Young, 2013).

Crucially, these stereotypical gender differences on interpersonal and agentic traits may be thought to mirror the two main dimensions found in facial first impressions, since the first facial dimension, trustworthiness, can be seen as primarily indexing intentionality, and the second dimension, dominance, primarily indexes capability (Oosterhof & Todorov, 2008). One possibility could be that male and female faces lie in the same space, with males represented as higher on the dominance dimension and females represented as higher on the trustworthiness dimension. Indeed, Sutherland and colleagues (in press) found that male faces are rated as higher on dominance but lower on trustworthiness than female faces on average. This parallels dimensional models of group perception in social psychology, such as the stereotype content model of person perception, which incorporates these perceived gender differences within a larger framework that describes the content of many stereotypes as having two key dimensions of warmth and competence (Cuddy et al., 2008; Fiske et al., 2007). In the stereotype content model, stereotypical women are located higher than men on the warmth dimension, and stereotypical men are located higher than women on the competence dimension (Cuddy et al., 2008).

However, these facial group differences might themselves be partially creating the dimensions of facial first impressions. First, cues to facial femininity and masculinity clearly underlie both dimensions, especially the dominance dimension (Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013). Second, we note that there is some evidence already for separate male and female dimensions for identity perception from faces (Jaquet & Rhodes, 2008). Moreover, we have previously found a third, youthful-attractiveness dimension for facial first impressions (Sutherland et al., 2013), and some studies have found differences in how attractiveness is perceived in male and female

faces (see Little et al., 2011; Rhodes, 2006 for reviews). It is therefore not clear what happens to these facial impression dimensions when the gender groups are split, but another possibility is that male and female faces lie in separate spaces in terms of social impressions.

Here we set out to establish which key dimensions underly facial first impressions for male and female faces and to ascertain the comparability of facial first impression dimensions across gender. As far as we are aware, the dimensional structure of traits has not yet been separately examined by face gender. This is theoretically important, because the evidence reviewed previously indicates that stereotypes of men and women are very different, giving potential reasons why dimensions of facial first impressions might also differ across face gender. While our main aim to improve the theoretical breadth of current models, establishing which dimensions subserve impressions to male and female faces could also have the benefit of improving computer models of facial perceptions to create more realistic computer generated faces (see Todorov et al., 2013; Walker, Jiang, Vetter, & Sczesny, 2011).

3.2.3. Dimensional accounts of social group and person perception

In parallel to the main aim of understanding how facial first impressions are organised for male and female faces, a secondary aim was to examine how male and female faces are perceived in terms of the stereotype content model's dimensions of warmth and competence. As described previously, men are rated as higher in competence and lower in warmth than women when verbal stimuli are used (Cuddy et al., 2008). Interestingly, the dimensions of warmth and competence also seemingly show a strong theoretical parallel to the trustworthiness (cf. warmth) and dominance (cg. competence) dimensions of facial first impressions described previously (Oosterhof & Todorov, 2008; Sutherland et al., 2013). For example, warmth and competence can also be considered to represent threat appraisal, with the first, warmth dimension corresponding to a judgement of the perceived intent of social groups (as cooperative or competitive), and the second, competence dimension forming a judgement of the perceived ability of that group to carry out these intentions, based on their power and social status (e.g. Fiske et al., 2007).

Given these potential theoretical similarities, the current research also attempted to understand how the facial first impression dimensions (trustworthiness and dominance) and the stereotype content model dimensions (warmth and competence) aligned, again by focusing on impressions made to female and male faces. In this way we aimed to uncover whether the dimensions of warmth and competence derived from the stereotype content model might help explain any facial gender differences found, since these models have been previously used to outline differences in how men and women are perceived (Cuddy et al., 2008).

3.2.4. Overview of current studies

In Study 1, we therefore examined whether the dimensions of facial first impressions are similar across face gender by exploring the factor structure of multiple social judgements for male and female faces separately. We then used confirmatory factor analysis to test the equivalence of dimensions across face gender. In Study 2, we followed up on our finding that the dominance dimension was not equivalent across gender by examining relationships between the two main dimensions of face perception (trustworthiness and dominance) with the two main dimensions of social group perception (warmth and competence) for male and female faces separately. The key finding was that dominance and competence judgements were less similar than warmth and trustworthiness judgements, especially for female faces. In Study 3, we showed that this pattern was robust across participants, and that male and female participants made very similar trait judgements. Finally, in Study 4 we replicated these findings using a more direct measure of similarity ratings between the members of pairs of average face images, either depicting trustworthiness and warmth, or dominance and competence.

3.3. Study 1

In Study 1, our principal aim was to assess the similarity of the structure of a wide range of first impressions to male and female faces, since gender-specific dimensional facial first impression models have not yet been examined. We expected that if the trustworthiness/valence and dominance dimensions are indeed by-products of fundamental evolutionary adaptations related to threat, as theorised (Oosterhof & Todorov, 2008), then they should appear for both male and female faces. A secondary

aim was to examine how the dimensions of the influential stereotype content model, warmth and competence, fit into facial first impressions models; since these stereotype content dimensions describe the evaluation of social groups including gender (Fiske et al., 2007).

We started with an exploratory factor analysis to examine the structure underlying a wide range of trait ratings made to 500 male and 500 female faces separately. In order to quantify where face gender differences lay, we then used confirmatory factor analyses to test each model (established on one half of the data) on the other half of the data; first by assessing whether the overall male and overall female models fitted better to the male or female data, and then by examining each model component in an iterative process to pinpoint specific differences. We predicted that a first, trustworthiness dimension would appear for both male and female impressions and that this dimension would be the most similar across facial gender given previous findings indicating that this dimension is primary in social judgements of others (Abele & Bruckmüller, 2011; Abele & Wojciszke, 2007). We left our expectations for other potential dimensions open, given that previous studies have found dimensions of trustworthiness, dominance and youthful-attractiveness of facial first impressions for male and female faces together (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Vernon et al., 2014) and that dominance, age and attractiveness can be evaluated differently for men and women (Little et al., 2011; Rudman, Moss-Racusin, Glick, et al., 2012; Sadalla, Kenrick, & Vershure, 1987; Sutherland et al., in press).

3.3.1. Ambient image faces

The stimuli used in Study 1 were a set of 500 female and 500 male naturalistic, highly varied "ambient image" face photographs which have been used in previous studies (Santos & Young, 2005, 2008, 2011; Sutherland et al., 2013; Vernon et al., 2014). Burton, Jenkins and Schweinberger (2011) first coined the phrase "ambient images" to represent the variability between images of faces we see in everyday life, and to stress the importance of including this variability (see also Jenkins et al., 2011). Following this approach, faces in this ambient image database are therefore intentionally chosen to vary on many potential cues to impressions including emotional expression, pose, lighting and other background characteristics, and facial paraphernalia (e.g. facial hair,

make-up, glasses: Santos & Young, 2005, 2008; Sutherland et al., 2013). However, the faces only depict adults of Caucasian appearance, since cross-cultural or own-race biases were not the focus of the current studies. The photographs in this database have been cropped around the head and shoulders and are standardised to be 150 pixels in height (approx. 5 cm on screen), but vary in width to preserve aspect ratio.

3.3.2. Methods 1

Eighteen traits were chosen to be included in the factor analyses for theoretical reasons in order that we could fairly test whether the dimensions identified by previous work were present for male and female faces separately. Approachability, trustworthiness, and smiling were selected as these are key traits or cues underlying the first facial impressions dimension identified by previous work (Oosterhof & Todorov, 2008; Sutherland et al., 2013); dominance, intelligence and sexual dimorphism (i.e. femininity-masculinity) were selected to represent the second dimension (Oosterhof & Todorov, 2008; Sutherland et al., 2013); and age, health, and attractiveness to index the third dimension reported in Sutherland et al. (2013). Ratings of warmth and competence were included to index the dimensions of the stereotype content model (Cuddy et al., 2008; Fiske et al., 2007). Finally, ratings of alternative aspects of face perception or facial first impressions were added to test whether the two dimensional models could account for these alternative judgements. Ratings of the arousal and pleasantness of targets' expressions were included to index Russell's (1980) circumplex model of affect, and ratings of babyfacedness and facial adiposity were added based on previous work indicating that these are important cues to first impressions (D. S. Berry & McArthur, 1986; Coetzee, Re, Perrett, Tiddeman, & Xiao, 2011).

Ratings of dominance, trustworthiness, approachability, degree of smiling, attractiveness, intelligence, sexual dimorphism, skin tone, confidence, aggressiveness, age, healthiness, facial adiposity, pleasantness of emotional expression, arousal of emotional expression, and babyfacedness for the 1,000 faces were taken from previous work (Santos & Young, 2005, 2008, 2011; Sutherland et al., 2013; Vernon et al., 2014). Ratings of warmth and competence were collected in the present study: ten participants (five male, mean age: 21.8 years) rated the 1,000 faces on their perceived warmth (1-7, with 1 being anchored as very cold, and 7 as very warm) and ten different participants

(five female, mean age: 21.0 years) rated the 1,000 faces on their perceived competence (1-7, with 1 being very incompetent, and 7 being very competent). Participants volunteered to take part in this study in return for course credit or a small remuneration and provided informed consent to procedures that were approved by the ethics committee of the University of York Psychology Department. All participants self-identified as culturally Western.

For all ratings, on each trial, participants saw one photograph with a Likert scale (1–7) presented underneath. Participants pressed the number key that corresponded with their rating and the next face photograph, randomly selected, appeared after a blank interval of approximately 750ms. Participants were given as much time as they wanted to make their rating, but were encouraged to go with their 'gut instinct' (Todorov et al., 2005). There were 10 practice trials with faces randomly selected from the database at the start. The inter-rater reliabilities for warmth and competence were good (Cronbach's alphas of $\alpha = .93$ and $\alpha = .72$ respectively) as were all previously collected traits (all alphas above .7: Sutherland et al., 2013; Vernon et al., 2014).

3.3.2.1. Exploratory factor analysis

In order to investigate the structure of first impressions made to male and female faces, exploratory factor analyses were run on the 18 ratings made to the 500 male and 500 female faces separately. Bartlett's test of sphericity indicated that the correlations were large enough that a factor analysis was appropriate: $X^2_{female}(153) = 9593.78$, $p < .001$; $X^2_{male}(153) = 9684.87$, $p < .001$. Kaiser's criterion and scree test, a parallel analysis and minimum average partial analysis were carried out in order to determine the number of factors (as in Sutherland et al. 2013; syntax from O'Connor 2000). For the female faces, the scree test, the MAP test and the parallel analysis returned three factors and Kaiser's criterion returned four factors, thus three factors were rotated. For the male faces, the scree test and the MAP test returned three factors and Kaiser's criterion and the parallel analysis returned four factors. A four-factor solution could not be rotated due to Heywood cases (i.e. the solution was not stable), thus three male factors were rotated. A principal axis factor analysis using direct oblimin rotation was used to determine the factor structure by interpreting the structure matrix and ignoring loadings below .3 (following Kline, 1994).

Table 3.1. Principal axis factor analysis: structure matrices for female and male faces (direct oblimin rotation). These can be interpreted as correlations between the factors and variables. Trait loadings above .3 are highlighted in bold.

	Factor 1		Factor 2		Factor 3	
n = 500 each	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
Approachability	**.95**	**.94**	.25	.29	-.12	.15
Aggressiveness	**-.92**	**-.91**	-.14	-.18	.13	-.02
Arousal	**.86**	**.87**	.22	.17	-.13	.11
Pleasantness	**.96**	**.96**	.22	.16	-.19	.21
Smiling	**.96**	**.96**	.15	.12	-.09	.15
Trustworthiness	**.85**	**.83**	**.42**	**.47**	-.19	.06
Warmth	**.97**	**.95**	.14	.22	-.10	.03
Competence	.18	.26	**.87**	**.79**	-.11	-.06
Confidence	**.47**	**.64**	**.66**	**.61**	-.26	.16
Dominance	**-.37**	-.15	.12	**.70**	**.59**	-.30
Intelligence	**.34**	.25	**.61**	**.69**	.02	-.09
Sex. dimorphism	-.29	.03	**-.43**	**.65**	**.59**	**-.48**
Skintone	.10	.17	.18	**.51**	.02	.00
Age	.06	.05	-.16	**.43**	**.82**	**-.92**
Attractiveness	.19	**.39**	**.76**	**.55**	**-.71**	**.67**
Babyfaced	.12	.27	.03	-.13	**-.72**	**.33**
Health	.17	**.35**	**.72**	**.57**	**-.69**	**.67**
Facial adiposity	.08	.06	**-.51**	.11	.14	**-.32**
Scale reliability α (traits loading ≤ .3)	.94	.94	.86	.80	.85	.69

The first factor was highly similar across face gender and seemed to index approachability (**see table 3.1.)**. Both trustworthiness and warmth loaded strongly on this approachability factor, which clearly approximated the first dimension in both facial (trustworthiness) and social group (warmth) models. However, the second factor was somewhat different across face gender. For male faces, it seemed best characterised as dominance/competence, with both of these traits loading highly. For female faces, dominance did not contribute highly to the second factor. Instead, this factor appeared only to index competence. The females did not show a clear dominance factor, as dominance ratings loaded largely on the other two factors (**see table 3.1.**). Interestingly, sexual dimorphism loaded negatively on the female competence factor, so that increasing masculinity decreased perceived competence (contrary to previous overall facial models: Oosterhof & Todorov, 2008; Sutherland et al., 2013). Finally, the third factor seemed similar in males and females, albeit in opposite directions (decreasing age/fitness for female faces, youthful-attractiveness for male faces). Since the sign of

the factor is arbitrary in factor analysis, further discussion focuses on both factors as youthful-attractiveness for simplicity. For a comparable orthogonal analysis (varimax), the variance explained by each female face factor was 34.9%, 17.2% and 15.6%; and for each male face factor, 35.8%, 19.9% and 12.8% respectively.

In order to visualise these factors (**see figure 3.1.**), we followed the procedure used by Sutherland et al. (2013) to create face-like averages from faces that lay high or low on each factor, measured through factor scores (using the regression method). The software Psychomorph was used to construct these averages (Tiddeman et al., 2001; for a practical guide to Psychomorph see Sutherland, 2015). As in Sutherland et al. (2013), ten faces were removed either because on closer inspection they seemed to be of another race, a celebrity, or they were very difficult to delineate due to head pose.

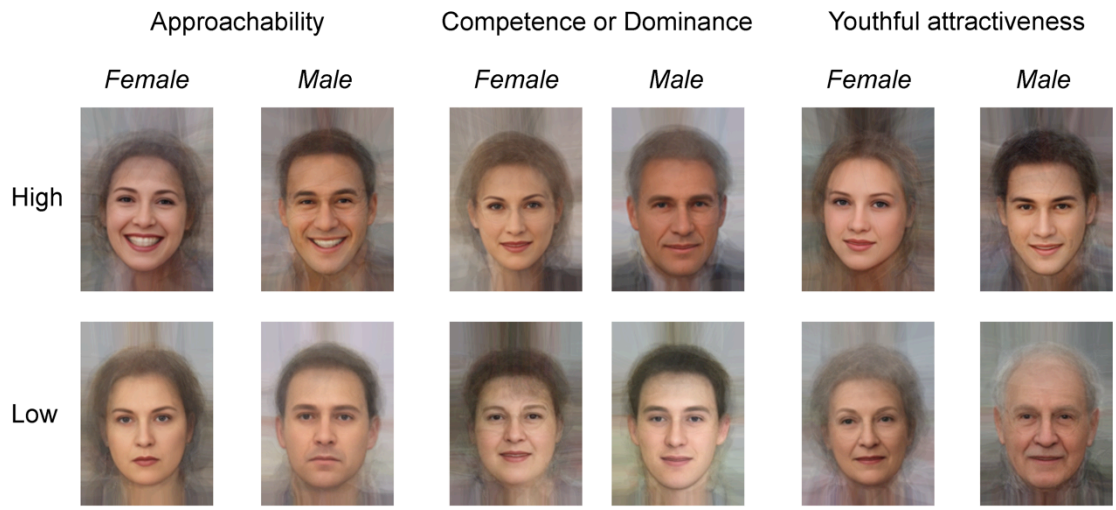

Figure 3.1. Face averages made from the 20 faces scoring highest and lowest for each factor identified through separate analyses for male and female faces. Factor scores were generated using the regression method. The female third factor (decreasing age/fitness) is transposed here as youthful-attractiveness for ease of comparison with the male factor (factor direction is arbitrary).

On the first factor, both the female and male high average face-like images (henceforward referred to as "faces" for simplicity) are clearly smiling, while the low male and female average faces are not (**figure 3.1.**). On the second factor, the high male average looks similar to the overall high dominance face average from Sutherland et al. 2013; **figure 2.2.** in this thesis), while the low face looks somewhat similar to the low

dominance face from Sutherland et al. (2013; **figure 2.2.** in this thesis). However, the female second factor high and low average faces clearly differ on cues such as attractiveness and femininity as well as competence or intelligence (**figure 3.1.**; compare these to the intelligence, attractiveness and femininity average faces from Sutherland et al. (2013; **figure 2.1.** in this thesis). Finally, both the female and male high average face on the third factor look younger and more attractive than the low average faces on this factor (**figure 3.1.**).

Finally, an analysis restricted to two factors was carried out to investigate how the structure of trait judgements for male and female faces related to previous two-dimensional facial first impression models (Oosterhof & Todorov, 2008; Walker & Vetter, 2009). For male faces, there were approachability and dominance/competence factors. For female faces, there were approachability and youthful attractiveness factors. This supports the suggestion that the first dimension (approachability, including warmth and trustworthiness) is the primary dimension in judging both male and female targets (e.g. Abele & Wojciszke, 2007; Fiske et al., 2007), but also further highlights key differences between male and female face models.

3.3.2.2. Confirmatory factor analyses

We first replicated the exploratory three-factor solution (principal axis factor analysis, direct oblimin rotation) on the first half of the data: initially for both male and female faces together (overall), then for female faces only, and finally for male faces only. To build the overall, female and male models, factor loadings were included and tested if they equalled .3 or above in the overall, female only or male only factor analysis from the first half of the data. To aid comparison, across models, we always scaled the factors to traits that seemed most robust across gender (factor one scaled to trait approachability, factor two to trait competence, and factor three to trait age; these traits were later also tested, see below).

We then used confirmatory factor analysis to test these three models on the ratings given to the other half of the faces, using the maximum likelihood method (T. A. Brown, 2006). First, overall, female and male face specific models were directly compared against each other on the male and female face data (**see table 3.2.**). Multiple indices

assessing different aspects of model fit were used to assess the relative fit between the three different models; including chi2, the Root Mean Square Error of Approximation (RMSEA), a comparative fit index (CFI), and a predictive fit index (Akaike's information criterion, AIC) following previous recommendations (T. A. Brown, 2006; Harrington, 2008). This showed that for ratings given to male faces, the overall model was the best fitting on all four indices, followed closely by the male-specific model. For ratings given to female faces, the female model was the best fitting on all four indices. Importantly, for both male and female faces, the opposite face-gender derived model was the worst fitting model (on all four indices).

Table 3.2. Comparison of overall, male face and female face derived models (data split half 1) with male and female face only data (data split half 2).

	Model	**Chi^2**	**df**	***p***	**CFI**	**RMSEA**	***p***	**AIC**	**Best model**
Male	Overall	826.92	125	<.001	0.849	0.15	<.001	918.92	*
faces	Female	1005.90	125	<.001	0.810	0.17	<.001	1097.90	
	Male	855.01	122	<.001	0.842	0.16	<.001	953.01	
Female	Overall	887.87	125	<.001	0.835	0.16	<.001	979.87	
faces	Female	806.793	125	<.001	0.853	0.15	<.001	898.79	*
	Male	1025.98	122	<.001	0.805	0.17	<.001	1123.98	

To investigate where these differences lie, a forward testing approach was chosen (following Byrne, 2004). First, we started with the most general model possible (i.e. a model including all factor loadings ≥ .3 from all three models; dimensions scaled as before), and fitted this general model to the male and female face sets simultaneously, allowing the male and female groups to freely vary. Parameters were then successively constrained to be equal across face gender groups, and chi^2 tests were used to test whether the new constrained model fitted significantly less well than the previous accepted model, in an iterative process. Importantly, these tests showed that factor 1 (approachability) did not significantly differ across male and female face groups (p = .081). However, factor 2 (dominance) and factor 3 (youthful attractiveness) were significantly different (both $p < .001$). For the dominance factor, the loadings of dominance, confidence, sexual dimorphism, skin tone, intelligence and also facial adiposity were significantly different across male and female face groups (all $p < .05$).

This indicates that this dimension differs on key traits across gender. For the youthful-attractiveness factor, only the facial adiposity, babyfacedness and dominance loadings were significantly different on the youthful-attractiveness factor (all $p < .001$), indicating that this factor is largely similar across face gender, with some sexually dimorphic attractiveness cues.

3.3.3. Discussion 1

Separate factor analyses of a wide range of facial trait judgements found a three-factor structure for both male and female ambient image faces. For male faces there were approachability, dominance/competence, and youthful attractiveness factors. For female faces, there were approachability, competence, and youthful attractiveness factors. The first approachability factor was clearly similar to previous social and facial dimensions of warmth and trustworthiness (Fiske et al., 2007; Oosterhof & Todorov, 2008) and was invariant across male and female faces. However, the competence/dominance and youthful attractiveness factors and their covariance were significantly different across male and female faces. Most strikingly, key aspects of the dominance/competence factor including dominance and sexual dimorphism (Oosterhof & Todorov, 2008; Sutherland et al., 2013) were not equivalent for male and female faces, with female faces lacking a strong dominance dimension. This suggests that models of facial first impressions need to take the gender of the face into account.

That female faces show a competence but not a dominance factor, unlike male faces, is also especially interesting given that the dimensions of warmth and competence form the basis for an influential model describing the content of social stereotypes (Cuddy et al., 2008; Fiske et al., 2007). Indeed, the theoretical underpinnings of dimensions of warmth and competence from models of inter-group perception seem to be highly similar to trustworthiness (cf. warmth) and dominance (cf. competence) models of facial first impressions. Our finding of gender differences seems to indicate that this similarity may mainly involve the first dimension, with the second dimension appearing to be relatively distinct for female and male faces. In Studies 2-4, we followed up this striking facial gender difference by focusing on trustworthiness and warmth, and dominance and competence judgements of male and female faces to establish how similar models of

facial impressions (Oosterhof & Todorov, 2008) are to models of social group perception (Cuddy et al., 2008) in the domain of facial first impressions.

3.4. Study 2

Our aim in Study 2 was to directly examine whether the two main dimensions derived from facial and social group studies are equivalent for judgements made from male and female faces. To the best of our knowledge, this has not yet been directly tested. In order to test this, we focused on trustworthiness and dominance judgements to represent facial dimensions, based on Oosterhof and Todorov (2008; and warmth and competence judgements to represent social dimensions, based on Fiske and colleagues' stereotype content model of social groups (Fiske et al., 2007). We tested the similarity between judgements of facial trustworthiness and facial warmth, and between facial dominance and competence, using correlations between ratings collected on the 1,000 ambient face images database. Given our previous finding of differences across facial gender, we examined impressions separately by facial gender.

3.4.1. Method and Results 2

Ratings of trustworthiness, warmth, dominance and competence from Study 1 were re-analysed here. In order to assess the claim that trustworthiness and warmth are highly similar trait judgements, we correlated the warmth and trustworthiness ratings from Study 1 (Sutherland et al., 2013) for the 500 male and 500 female face photographs separately (**see figure 3.2.**). The correlation between trustworthiness and warmth was substantial and highly significant for male faces: $r_{male} = .82$, $p < .001$; and female faces: $r_{female} = .81$, $p < .001$. There was no difference between the size of these correlations across face gender ($Z = -0.23$, $p = .82$). It was therefore clear that trustworthiness and warmth were highly similar judgements for both male and female faces.

In order to assess the claim that dominance and competence are highly similar trait judgements, we correlated the competence and dominance ratings from Study 1 (Sutherland et al., 2013) for the 500 male and 500 female face photographs separately (**see figure 3.2.**). The correlation for dominance and competence was moderate and significant for male faces: $r_{male} = .50$, $p < .001$; but, although statistically significant,

was small for female faces: $r_{female} = .12$, $p < .01$. The size of the correlation between dominance and competence was significantly greater for male faces than female faces ($Z = -6.63$, $p < .001$). It therefore appears that dominance and competence judgements are not particularly closely related for female faces. In contrast, dominance and competence ratings are similar, although far from identical, for male faces (**see figure 3.2.**).

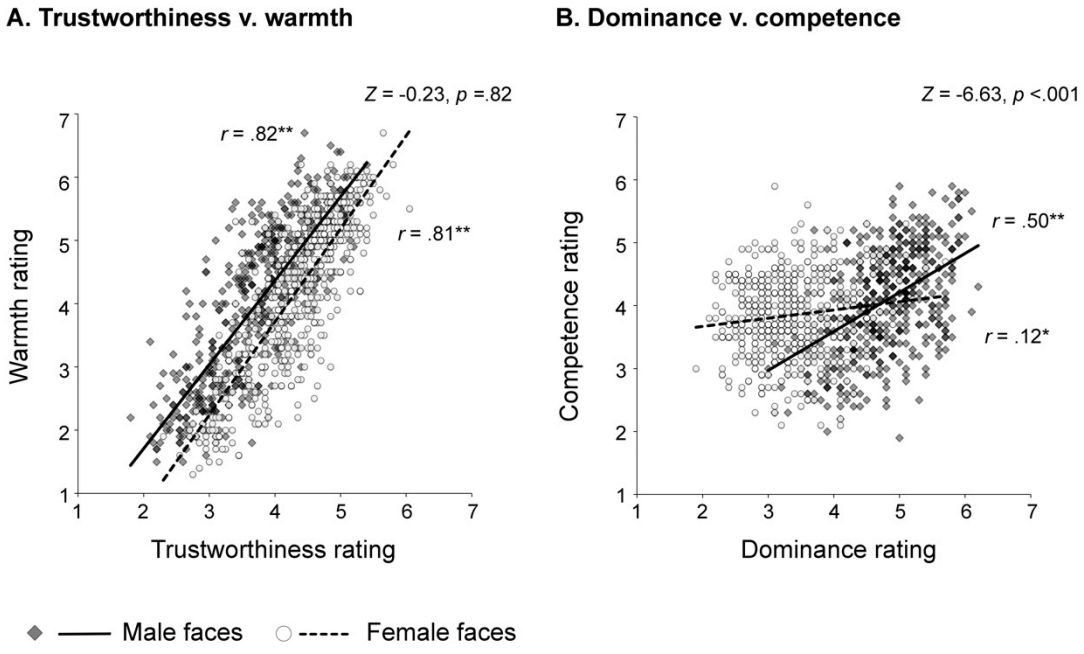

Figure 3.2. The correlation between A) average trustworthiness and warmth judgements; and B) average competence and dominance judgements for male and female face photographs separately. ** $p < .001$, * $p < .05$.

3.4.2. Discussion 2

In Study 2 we found that while trustworthiness and warmth judgements were highly related for both male and female faces, dominance and competence judgements were only moderately related for male faces and only weakly related for female faces. This pattern is consistent with the idea that the first dimension described in facial first impressions models (e.g. trustworthiness: Oosterhof & Todorov, 2008) is the same as the first dimension found in social psychology models of inter-group perception (e.g. warmth: Fiske et al., 2007). However, our findings show that the second, dominance

dimension found in face perception models (Oosterhof & Todorov, 2008) is not the same as the competence dimension found in social psychology (e.g. Fiske et al., 2007, at least for female faces. This helps narrow the locus of the facial gender differences found in Study 1.

Since Study 2 did not set out to examine participant gender, the possibility remains that the judgements of participants of one gender might be responsible for this effect. Moreover, it would be useful to know whether males and female raters generally agree in their judgements of faces, since this would allow future research to include or ignore participant gender as a factor. Finally, Study 2 looked at correlations for aggregated data generated by averaging across responses from individual participants to each face photograph. While this approach was justified by our focus on the judgements that people make in general, and by the high agreement between raters, we considered it useful to ascertain whether our findings in Study 2 were robust across participants.

3.5. Study 3

In Study 3, our main aim was to determine whether our finding from Study 2 that facial competence and dominance are less similar than facial warmth and trustworthiness was reliable across participants, and whether or not participant gender had any major influence on this finding. Since studies consistently report high interrater agreement with mixed-gender participant groups (e.g. Oosterhof & Todorov, 2008; Sutherland et al., 2013), we predicted that male and female participants' trait ratings of faces would substantially agree and that ratings would not be significantly different across own-gender or other-gender groups. Based on Study 2, we also expected to find higher agreement (correlations) between warmth and trustworthiness judgements than competence and dominance judgements across participants; and we also expected that this would interact with facial gender as before.

To set our findings in a broader context, we were also curious as to whether male and female raters would agree on trait perceptions more widely, since this has not been systematically investigated. We therefore also selected additional traits from previous research as a comparison, including approachability, smiling, intelligence, sexual

dimorphism, age, babyfacedness, health, and attractiveness (D. S. Berry & McArthur, 1986; Cuddy et al., 2008; Oosterhof & Todorov, 2008; Sutherland et al., 2013).

3.5.1. Methods 3

Twenty-four participants (12 female; mean age: 20.6 years) volunteered to take part in the second study in return for course credit or a small remuneration. Two additional participants only took part in the first session: their data was excluded. Participants provided informed consent to procedures that were approved by the ethics committee of the University of York Psychology Department. All participants self-identified as culturally Western. Participants did not take part in the other currently reported experiments.

3.5.1.1. Stimuli & Procedure 3

Two hundred face stimuli (100 female) were chosen at random from the overall 1,000-strong ambient image dataset. Participants were tested in a quiet room on a PC running E-Prime software (version 2; Psychology Software Tools, Pittsburgh, USA). Participants rated all 200 faces on 12 different traits, with each trait rated in a separate block of 200 trials with 24 practice faces (not analysed) presented before each new block. We included trustworthiness, warmth, dominance and competence, as key traits relevant to the current paper. For a broader background against which to evaluate results for the four key traits, we also chose additional ratings of approachability, attractiveness, age, babyfacedness, intelligence, health, smiling, and sexual dimorphism to sample a wide variety of traits based on previous work (D. S. Berry & McArthur, 1986; Cuddy et al., 2008; Oosterhof & Todorov, 2008; Sutherland et al., 2013).

Participants rated all traits during two sessions in an order that was counterbalanced across participants in such a way that one participant of each gender carried out the ratings in the same order (these were ‘matched’ gender pairs). However, we note that this procedure of having many multiple traits rated will almost certainly maximise carryover effects, where the first trait ratings artificially influence the second trait ratings, thus inflating inter-trait correlations (i.e. Rhodes, 2006). Since our experimental aims would be affected by inflated correlations and by judgements on one trait priming judgements on another, we refrained from calculating correlations between ratings made

by the same individual, but instead used a Leave-One-Participant-Out analysis to compare each individual's judgements with the average judgement across the rest of the group (see **Analysis 3**).

3.5.1.2. Analysis 3

In order to establish how well male and female participants agreed with each other, we carried out Leave-One-Participant-Out (LOPO) analyses with a technique similar to that used by Zebrowitz and colleagues (2013). In a LOPO analysis, the pattern of response for an individual participant is compared to the pattern for the rest of a participant group. Here, individual participants' responses were correlated to the mean for other participants of the same gender, minus their own response (within-gender correlations) or to the mean for participants of the opposite gender (between-gender correlations). Within-gender and between-gender correlations could then be compared across participants to test whether participant gender affected the size of the correlations (see Zebrowitz et al., 2013 for a similar analysis). If participants differed in their perception of traits by gender, then the within-gender correlations would be significantly higher than the between-gender correlations.

To test whether our previous findings (that dominance is less highly correlated with competence than warmth is with trustworthiness, especially for female faces) were consistent across participants, we employed similar LOPO analyses by correlating each individual participant's rating on one trait on a dimension (e.g. dominance or trustworthiness) with the mean of the rest of the group's ratings on the proposed equivalent trait (e.g. competence or warmth) for male and female faces. This allowed us to compare the consistency of judgements across facial dimensions and face gender, to assess whether our findings from Study 2 were consistent across participants, while avoiding confounding carryover or priming effects. Participant gender was also entered as a factor.

3.5.2. Results 3

3.5.2.1. Reliabilities

Reliability coefficients (Cronbach's alpha) across all participants were high, which itself demonstrates substantial interrater agreement across participant gender. All reliabilities

were above .7 for all faces, and for male and female faces separately. Agreement was also high between male and female participants' average ratings of the twelve traits for both male and female faces **(see table 3.3.).** Again, this indicates that there are no major differences in the ratings of male and female perceivers.

Table 3.3. Correlations between male and female participants' ratings for all faces, female faces, and male faces.

		All faces	Female faces	Male faces
Key traits	Trustworthiness	.87**	.83**	.87**
	Warmth	.92**	.91**	.93**
	Competence	.77**	.67**	.84**
	Dominance	.84**	.75**	.87**
Additional	Age	.97**	.98**	.98**
traits	Approachability	.90**	.89**	.90**
	Attractiveness	.89**	.91**	.83**
	Babyfacedness	.79**	.80**	.81**
	Health	.88**	.87**	.90**
	Intelligence	.83**	.74**	.88**
	Sex. dimorphism	.97**	.87**	.82**
	Smiling	.98**	.98**	.97**

** all $p < .001$, $N = 200$, female faces $n = 100$, male faces $n = 100$

3.5.2.2. Background: participant gender

In order to statistically test male and female agreement in first impressions, we carried out a Leave-One-Participant-Out (LOPO) analysis by correlating each participant's ratings on a given trait with the average of all other within- or between-gender ratings for this trait, excluding the participant or their matched opposite-gender partner (see also Zebrowitz et al., 2013). Since the genders were matched by the order in which they rated the traits, removing the matched opposite sex partner stopped this matching from artificially increasing agreement. These LOPO analyses were carried out for male and female faces separately and for each of the 12 traits. Correlations were *z*-transformed before analysis using Fisher's *r*-to-*z* transform and all analyses were corrected using the Greenhouse-Geisser correction for sphericity where applicable. This method allows us

to examine the effect of rater gender (between-gender or within-gender), face gender (male or female) and trait, as well as interactions between these factors.

The overall three-way interaction was not significant: $F(11,253) = 0.93$, $p = .512$. Importantly, neither of the interactions between participant gender group (between-gender or within-gender) were significant: participant gender with face gender (male/female): $F(1,23) = 1.74$, $p = .20$; participant gender and trait: $F(3.82,87.94) = 1.67$, $p = .166$. However, contrary to our predictions, participants agreed very slightly but significantly more with other participants of their own gender than the opposite gender (mean difference of $r = .01$; significant main effect of gender of participant: $F(1,23) = 5.21$, $p = .032$. There was also a main effect of trait, with some traits showing higher agreement across all participants: $F(4.71, 108.22) = 86.82$, $p < .001$. The face sex main effect was not significant: $F(1,23) = 2.56$, $p = .123$. However, the interaction between face gender and rated trait was highly significant: $F(5.71,131.35) = 23.32$, $p < .001$. Again, this suggests that the main differences in facial first impressions lay between male and female faces rather than perceivers.

Table 3.4. Mean correlation coefficients between each individual's ratings with the rest of the group ($n = 24$), separately for the twelve rated traits and face gender. Traits are grouped in terms of differences between male and female faces.

		Male faces	**Female faces**	**Statistical significance of difference**
Male faces >	Competence	.55	.41	$p < .001$
Female faces	Dominance	.57	.40	$p < .001$
	Intelligence	.62	.47	$p < .001$
Male faces =	Approachability	.65	.64	$p = .947$
Female faces	Age	.87	.87	$p = .269$
	Babyfacedness	.56	.53	$p = .171$
	Health	.59	.59	$p = .777$
	Trustworthiness	.55	.52	$p = .225$
	Warmth	.69	.71	$p = .090$
Male faces <	Attractiveness	.57	.67	$p < .001$
Female faces	Sexual dimorphism	.48	.60	$p < .005$
	Smiling	.83	.86	$p < .005$

We followed up the face gender by trait interaction by comparing male and female faces for each trait across all raters, ignoring participant gender (between v. within) as a factor (see **table 3.4.**). Participants showed higher agreement for male than female faces for dominance, competence and intelligence, and higher agreement for female than male faces for smiling, attractiveness and sexual dimorphism (**see table 3.4.** all $p < .005$; differences still significant after Bonferroni correction). That some judgements show higher agreement for female faces indicates that female faces are not simply harder to evaluate. Individual trait correlations between each rater with the average of the rest of the group for male and female faces were all significantly higher than zero on average, indicating that all traits were judged with at least some agreement across raters (one-sample *t*-tests, $n = 24$ all $p < .001$, **see table 3.4.** for values).

3.5.2.3. Key traits: facial gender

We then sought to determine whether the data would replicate the finding in Studies 2 and 3 that warmth/trustworthiness judgements were more similar than dominance/competence judgments, especially for female faces. Since participants rated the same faces several times, we did not correlate participants' ratings on one trait with their rating on another; instead, we used the LOPO technique again. Here, each participant's ratings on one trait on a given dimension (e.g. dominance or trustworthiness) was correlated with the average of all other participants' ratings on the putative other trait on that dimension (e.g. competence or warmth). Correlations were transformed using Fisher's *r*-to-*z* transform before analysis.

Averaged LOPO correlations are shown in **figure 3.3.** There was a significant main effect of dimension, so that facial trait judgements on the proposed first dimension (trustworthiness with warmth) were more correlated than facial trait judgements on the proposed second dimension (dominance with competence): $F(1,23) = 111.71$, $p < .001$ replicating the overall difference seen in Studies 2 and 3 (**see figure 3.3.**). There was also a main effect of face gender: $F(1,23) = 30.15$, $p < .001$. Crucially, the interaction between the two dimensions and face gender was significant: $F(1,23) = 9.76$, $p < .01$. Pairwise comparisons indicated that the correlations between dominance and competence ratings were higher for male than female faces: $t(23) = 6.14$, $p < .001$; however, the correlations between ratings of warmth and trustworthiness did not differ

between male and female faces: $t(23) = .77$, $p = .447$. Thus, as in Studies 2 and 3, dominance is less highly correlated with competence than warmth is with trustworthiness, now across participants, and this is especially true for female faces **(see figure 3.3.**).

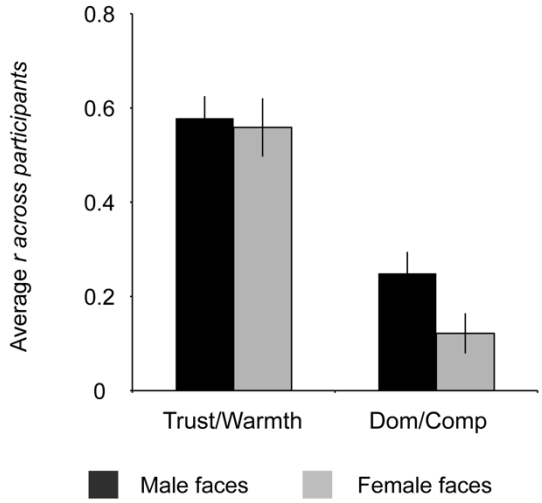

Figure 3.3. Average correlations for male and female faces between individual participants' ratings of trustworthiness/warmth with mean warmth/trustworthiness ratings for the rest of the group; and between individual participant's ratings of dominance/competence with mean competence/dominance ratings for the rest of the group. Error bars are 95% CI.

Finally, these analyses were re-run with participant gender as a factor (here: male or female). The four-way interaction with participant gender was not significant, thus the overall pattern of findings did not change across participant gender: $F(1,22) = .004$, $p = .95$.

3.5.3. Discussion 3

Study 3 found an interaction between the impression dimension (dominance/competence versus trustworthiness/warmth) and face gender. This replicates Study 2 by showing that dominance and competence judgements were less correlated across participants than warmth and trustworthiness judgements, especially for female faces. Again, this has implications for claims made about the similarity of face perception and social psychological models (see the **General discussion**).

We found that male and female perceivers showed considerable agreement in their facial first impressions, with high correlations between their group ratings, and high interrater reliabilities. However, we also found a small but significant main effect of perceiver gender, so that participants agreed slightly more with other participants of their own gender. Future studies should consider whether this is an issue worth controlling for, depending on the study aims and context. Importantly, here we did not find that participant gender qualified our overall findings.

In Study 4, we aimed to replicate the finding that dominance and competence judgements were less related than warmth and trustworthiness judgements, especially for female faces, using a very different and more direct measure of similarity.

3.6. Study 4

In Study 4, we asked participants directly to evaluate the similarity between same gender pairs of average images of faces of high or low warmth and trustworthiness, or high or low dominance and competence. In this way, we could directly examine how similar warmth and trustworthiness pairs of average male or female faces were perceived to be, compared to dominance and competence average face pairs. Using average face images also offered a well-controlled method of examining these differences, since only cues to trait judgements which are consistent across the individual faces remain in the face averages. Finally, asking for an evaluation of image similarity in all experimental conditions also avoids any ambiguities in language use.

We sought to use this novel method to replicate the main findings of Study 2 and 3. Based on these findings, we predicted that the dominance and competence average images would be seen as less similar to each other than the warmth and trustworthiness average images. We also predicted that there would be an interaction with face gender as before, so that this similarity difference between the dimensions would be greater for female than male faces.

3.6.1. Methods 4

Twenty-four participants (12 male; mean age: 23.2 years) volunteered to take part in the third study in return for course credit or a small remuneration. Participants provided electronic informed consent to procedures that were approved by the ethics committee of the University of York Psychology Department. The experiment was hosted online by Qualtrics (www.qualtrics.com) although participants were recruited through the University. All participants self-identified as culturally Western. Participants did not take part in the other currently reported experiments.

3.6.1.1. Stimuli & Procedure 4

Stimuli were created using Psychomorph (version 5, Tiddeman et al., 2001 by averaging the 20 ambient image faces rated highest and lowest on warmth, trustworthiness, competence and dominance (see Sutherland et al., 2013 for a full description of averaging procedures or Sutherland, 2015 for a practical guide to Psychomorph). The stimuli are depicted in **figure 3.4.** Stimuli were cropped around the face and aligned so that their eyes were horizontal in order to stop head tilt affecting the similarity judgements, since the pairs were to be presented directly side by side. As Study 1, nine faces were removed for technical reasons. By averaging across individual exemplars, the resulting face-like average images (referred to as 'faces' for simplicity) should represent the consistent cues underlying the perception of these traits. This also allowed us to examine the four key judgements directly and with different stimuli from those used in the previous three studies.

In a first block of trials, pairs of averaged faces were simultaneously presented and participants rated how similar the faces in each pair were (1 not very similar - 7 very similar). Eight pairs were used, which contrasted either dominance and competence average images or warmth and trustworthiness average images, for a given gender and at a given high or low extreme. For example, the low warmth male average image was compared to the low trustworthiness male average image, and so forth. Pairs were rated for their similarity three times, with the first set of ratings as practice trials (not analysed). Within the experiment, the order of the side of presentation of each of the average images in a pair was counterbalanced across trials.

In second and third blocks, each average image face was rated along the two model dimensions, with half of the participants rating the faces on the face perception model dimensions (trustworthiness and dominance, block order counterbalanced) and the other half on the social group model dimensions (warmth and competence, block order counterbalanced). Again, participants rated the faces three times in a block, with the first time as a practice (not analysed). These ratings were made to confirm that the average faces were indeed perceived as high or low on the manipulated trait, so that any lack of similarity between pairs of average faces could not be explained by a failure of the stimuli to represent these characteristics adequately.

3.6.2. Results 4

3.6.2.1. Similarity ratings

The second and third similarity ratings were averaged together. A three-way ANOVA was run (n = 24) on the similarity ratings with the dimension of the face (2 levels: trustworthiness/warmth, or dominance/competence), the dimensional position of the face (2 levels: high or low) and the gender of the face pair (female or male) as within-subjects factors. The three-way interaction was significant: $F(1,23) = 7.91$, $p = .010$; and, as predicted, there was a significant two-way interaction between the gender of the face and the dimension: $F(1,23) = 13.33$ $p < .001$; as well as between the dimension and the position of the face: $F(1,23) = 76.75$, $p < .001$. Also as predicted, there was a main effect of the dimension of the face, so that the trustworthiness/warmth pairs were rated as significantly more similar than the dominance/competence pairs: $F(1,23) = 217.13$, $p < .001$. There were also main effects of the average image position, so that face pairs high on a trait were rated as more similar than low face pairs: $F(1,23) = 8.68$, $p < .01$; and of face gender, so that male average face pairs were rated as more similar than female pairs: $F(1,23) = 19.44$, $p < .001$; **see figure 3.4.**).

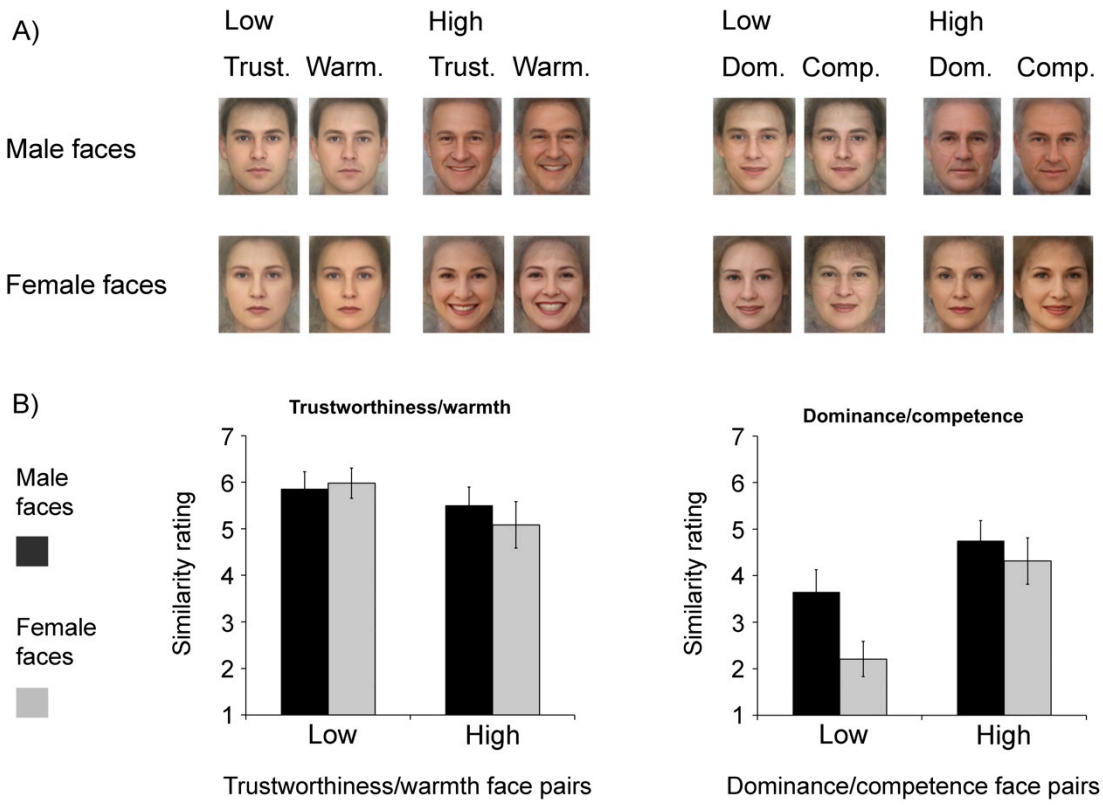

Figure 3.4. A) Pairs of male and female face-like averages made from the 20 face photographs rated highest and lowest for trustworthiness and warmth; or dominance and competence. B) Similarity ratings (1-7) for trustworthiness and warmth pairs; and dominance and competence pairs of male and female face-like averages. Error bars are 95% CIs.

Planned comparisons examined the similarity of average faces on each proposed dimension separately. For trustworthiness/warmth average face pairs, there was no main effect of gender: $F(1,23) = 0.82$, $p = .373$, and facial gender did not interact with the position of the face: $F(1,23) = 2.28$, $p = .144$. There was only a main effect of the position of the face, so that low average warmth or trustworthiness faces were seen as more similar than high average warmth or trustworthiness faces: $F(1,23) = 6.87$, $p = .015$. However, for the dominance/competence average face pairs there was the expected main effect of face gender; so that female competence and dominance average face pairs were seen as significantly less similar than male competence and dominance pairs: $F(1,23) = 31.49$, $p < .001$. This was moderated by a significant interaction between face gender and the position of the average image: $F(1,23) = 7.36$, $p = .012$. Planned comparisons showed that the low competence/dominance female average image pairs were rated as significantly less similar than the male average pairs: $t(23) =$

7.52, $p < .001$ (**see figure 3.4.**); but the high competence/dominance male and female pairs did not significantly differ: $t(23) = 1.48$, $p = .152$ (**see figure 3.4.**). There was also a main effect of the dimensional position of the face, so that high dominance and competence face averages were seen as more similar than low pairs: $F(1,23) = 83.61$, $p < .001$.

Finally, as expected, there were no main effects of participant sex, and nor did participant sex significantly interact with our current findings when entered as between-participant factor: all $p < .10$. The theoretically interesting three-way interaction between dimension, face gender and position; and the main effect of dimension were both still significant when participant sex was added (both $p < .01$).

3.6.2.2. Trait ratings

Table 3.5. Mean trait ratings (and standard deviations) for high and low male or female average faces, for each manipulated trait.

	Face gender	High average face	Low average face	High vs. low comparison
Trustworthiness faces/rating	Female	5.92 *(0.87)*	3.67 *(1.48)*	$p < .001$
	Male	5.29 *(1.18)*	3.38 *(1.65)*	$p = .011$
Warmth faces/rating	Female	6.04 *(1.08)*	2.42 *(1.51)*	$p < .001$
	Male	6.08 *(1.08)*	1.96 *(0.96)*	$p < .001$
Dominance faces/rating	Female	5.50 *(0.83)*	2.50 *(1.02)*	$p < .001$
	Male	5.33 *(0.54)*	3.17 *(0.91)*	$p < .001$
Competence faces/rating	Female	5.42 *(0.73)*	3.46 *(1.10)*	$p < .001$
	Male	5.71 *(1.01)*	3.67 *(0.98)*	$p < .001$

After discarding the first ratings as practice trials, the second and third trait ratings were averaged together for each participant, rated trait and stimulus face. The trait ratings showed that the pairs of averaged images (high and low on a given trait, within gender) differed on the manipulated trait, as expected ($n = 12$; all $p < .011$; **see table 3.5.**). All were still significant after Bonferroni correction except the comparison between the trustworthiness high and low male faces, although the means were in the expected

direction. Importantly, the male and female dominance and competence trait face averages were rated as expected on these traits; thus our results on the similarity measure were not merely due to a failure of the average images to represent the manipulated trait dimension.

3.6.3. Discussion 4

Study 4 found that the dominance and competence face averages were perceived as less similar to each other than the warmth and trustworthiness face averages. We also again found a facial gender difference: this difference between the dimensions was greater for female faces than male faces, as predicted. This conceptually replicates the overall pattern found in Studies 1-3, using a more direct measure of similarity. This pattern was qualified by a three-way interaction with dimensional position so that the low dominance and competence female faces were seen as significantly less similar to each other than the low dominance and competence male faces.

3.7. General discussion

Our principal aim in the current paper was to explore how facial first impressions are organised for male and female faces when examined separately. In Study 1, we found a three-factor structure for both male and female faces, with the first (approachability) factor being equivalent for female and male faces, but not the second (competence/dominance) or third (youthful-attractiveness) factors. In the three remaining studies, we focused on the first two dimensions due to their importance in previous facial and social models of impressions (Fiske et al., 2007; Oosterhof & Todorov, 2008). In Study 2, using trait ratings and real faces, we found that while trustworthiness and warmth ratings of 1,000 ambient image faces are highly related for both male and female faces, dominance and competence ratings of these faces are less strongly related, especially for female faces. In Study 3, we established that this pattern of face gender differences was robust across male and female participants. In Study 4, we replicated these dimensional differences across face gender using direct ratings of similarity and average face images. Our findings have implications for current dimensional models of facial impressions (Oosterhof & Todorov, 2008; Sutherland et al., 2013) and of group perception (Fiske et al., 2007), as discussed below.

3.7.1. Social and facial models of person perception

Our main finding was that there are categorical gender differences in the dimensions of facial first impressions, with female faces showing a competence rather than a dominance factor. This provides an important constraint to previous facial first impression models (Oosterhof & Todorov, 2008; Sutherland et al., 2013). Taken together, these findings also support the claim that the first dimension found in facial impressions research (trustworthiness: Oosterhof & Todorov, 2008) and the first dimension found in social psychology studies (warmth: Fiske et al., 2007) are very similar for judgements made from faces. However, our current findings indicate that the second dimension found in studies of face perception (dominance: Oosterhof & Todorov, 2008) is not identical to the second dimension found in social group research (competence: Fiske et al., 2007), especially for impressions of female faces. Here we will attempt to resolve the current findings with claims that the dimensions underlying social group and facial first impression models are similar (Montoya & Horton, 2014; Sutherland et al., 2013; Todorov, Baron, et al., 2008).

In considering this question, it is important to stress that the similarity between the face perception and social psychological models also rests on their theoretical underpinnings. Specifically, both Oosterhof and Todorov (2008) and Fiske and colleagues (Fiske et al., 2007) relate the first dimension to a person or a group's intentions, with the second dimension being their capability to carry out their intentions. Therefore, it is possible that the cues underlying capability differ across abstract (e.g. verbal) and facial stimuli. Potentially, visual facial stimuli offer more cues to physical capability than verbal group stimuli; for example, cues to physical dominance, rather than social competence or status as for groups (Fiske et al., 2002). This could also explain the lack of a clear female dominance dimension, because capability appraisal based on physical dominance may simply be less relevant for female targets. Human females have less upper body strength than males (Sell et al., 2009) and cues to dominance include physical strength and masculinity (Oosterhof & Todorov, 2008). Perceivers are in fact less accurate at identifying female than male strength from faces (Sell et al., 2009). Moreover, the use of different cues for judging the second dimension in male and female faces also fits with the finding that facially dominant men and women are

evaluated differently, so that dominance is evaluated negatively for female faces (Sutherland et al., in press). This account therefore resolves differences we noted between male and female faces, and between face and social models, by first highlighting that the same underlying theoretical emphasis on intentions and capability underlies both models; and second, by suggesting that cues to capability would be expected to vary depending on the stimuli or target group. This account moves the field forward in several important ways, which are outlined below.

3.7.2. Intentionality and capability

First, we highlight that both social psychological and facial impression models have an underlying theoretical framework that emphasises that proximal judgements are ultimately used to judge intentionality and capability. This moves the theoretical focus away from the semantics of specific traits, and onto the functionality of the processes of person evaluation. Todorov and colleagues (Todorov, Said, et al., 2008) have also argued that a focus on specific traits is unhelpful in a variety of research contexts, including in neuroimaging studies, because many traits are highly inter-correlated. To illustrate, if one finds that attractiveness and competence perceptions are both correlated with an outcome (e.g. votes cast; Verhulst, Lodge, & Lavine, 2010) then it becomes hard to evaluate the causal pathway linking the traits to these outcomes. Often the high multicollinearity between traits makes it impossible to experimentally manipulate specific traits orthogonally. This problem can be minimised by looking at broad dimensions of evaluation rather than specific traits (as suggested by Todorov and colleagues, 2008), but here we also point out that focusing on the theoretical processes behind specific traits can also help avoid these problems. By grounding their choice of traits in expected functionality, future researchers can focus on conceptually important differences between traits (this is returned to below). Similarly, a focus on functionality can help emphasise the important commonalities across different studies using different trait labels (e.g. trustworthiness or valence: Oosterhof & Todorov, 2008; versus approachability: Sutherland et al., 2013), or stimuli (e.g. male versus female faces, as in the current study).

3.7.3. The importance of the context

A second important implication arising from the finding that impressions of male and female faces are organised differently, is that this highlights the importance of the context around the face. Here, we argue that the gender of the face (with associated stereotypes) forms one such context, with clear differences arising from the cues available in male and female faces. We suggest that other such contexts will also be important. For example, our findings open up the possibility that the organisation of impressions of young and old faces (or own and other-age face perceptions) may also differ in structure. Likewise, we expect that dimensions may show culture-specific aspects when perceivers from different cultures (nationalities, ethnicities, social classes etc.) are making judgements, and that there may be differences in terms of how own- and other-culture faces are viewed.

Placing a focus on the underlying theorised motivations of intentionality and capability gives both facial first impressions models and models of person perception in general considerable flexibility in explaining how first impressions occur across a multitude of contexts. This can potentially help explain a broader range of results. For example, with this framework in mind, traits frequently inferred from faces by non-Western perceivers may not exactly translate to 'trustworthiness' or 'dominance' (and nor should we expect them to, given cultural or language differences), but nevertheless they should serve to elucidate a target's intentions towards the perceiver or a salient other, and their capability in carrying out these intentions.

To be clear, we are not arguing that researchers ignore or minimise the differences found across cultural and experimental contexts. On the contrary, we think that the next stage of research on facial and person evaluation should explicitly consider how cultural and experimental contexts, along with physical, occupational, goal-driven, and perceiver contexts, might shape target perceptions. In this respect, we join previous authors who have recently argued that a thorough theoretical understanding of face perception will only be achieved by a systematic programme of research which integrates top-down perceptions from the context with bottom-up cues in the face (Aviezer, Hassin, Bentin, & Trope, 2008; Quinn & Macrae, 2011; Todorov, Said, et al., 2008). In fact, suggesting that the underlying processes in person perception might

ultimately function to estimate targets' intentionality and capability explicitly recognises that different proximal judgements could be formed or cues utilised depending on the perceiver, target and location, and how these stimuli are presented to the perceiver. This opens the way to more specific and targeted analyses of how impressions of people might change across different contexts. Examples of more specific reasoning along these lines are given below.

3.7.4. Face and social dimensions

In the current study we found that trustworthiness and warmth judgements made from faces were highly related (Studies 1, 2 and 3) and that they loaded on the same factor (Study 4). This agrees with Abele and Wojciszke's (2007) similar finding for abstract judgements of these traits. However, this finding does not fit with studies showing that trustworthiness and warmth judgements can be dissociated in extra-facial person perception (Brambilla, Rusconi, Sacchi, & Cherubini, 2011; Goodwin, Piazza, & Rozin, 2014; Leach, Ellemers, & Barreto, 2007). This discrepancy might occur because the current study used facial stimuli and previous studies used verbal stimuli. Potentially, people might rely on a more rudimentary approach/avoid judgement for faces, based on salient cues since they lack information needed to disambiguate subtle differences. Alternatively, specific contextual manipulations might find situations where facial trustworthiness and warmth judgements diverge since they can be conceptually dissociated (Goodwin et al., 2014). In either case, we would suggest that both warmth and trustworthiness are still facets of a higher-order judgement of intentionality since both still refer to a target's interpersonal orientation towards the perceiver.

These differences notwithstanding, we also still expect that judgements on the second dimension would vary more by the extra-facial context. This is based both on our current findings and an analysis of the theoretical underpinning of this dimension, since by definition, being competent depends on the skills needed for the task at hand. Moreover, perceived facial competence but not dominance predicts the success of a target in political contexts (Chiao, Bowman, & Gill, 2008) whereas perceived facial dominance predicts success in the military (Mueller & Mazur, 1996). In some situations, task competence might even partially rely on being sociable or trustworthy. This could be assessed by examining the correlation between facial trait judgements

while manipulating different contexts. We would predict that cues underlying competence judgements would differ in a top-down way depending on competencies perceived as necessary, based on studies showing that perceivers use facial cues that they expect to be most relevant (Oldmeadow et al., 2013; Olivola & Todorov, 2010a). It would also be interesting in future to examine whether there are differences in the trait structure between visual and verbally presented target stimuli, or between perceptions of individuals and social groups. Finally, the local context of the experimental set-up may also change the saliency of social groups and this in turn may change the cues to used judge faces (e.g. similarly to shifting standards in stereotyping research: Biernat, 2012). For example, differences between male and female faces may be less strong if faces are viewed separately (cf. Kramer, Jones, & Sharma, 2013). Ultimately, we emphasise again that exploring how the context combines with visual cues in a face to form perceivers' first impressions is a necessary next step for social face perception research.

3.8. Conclusions

To the best of our knowledge, this is the first set of studies to directly test the equivalency of male and female dimensions of facial first impressions. This was also a first step in examining how facial (trustworthiness and dominance) and social group (warmth and competence) dimensions might relate to one another. We found that while trustworthiness and warmth dimensions were highly similar, dominance and competence dimensions were not highly related, especially for female faces. The current results highlight interesting differences between social psychological models of extra-facial person perception and facial models of first impressions. In particular, the current results speak to the need to integrate specific contexts (here, based on the social category of gender) with dimensional accounts of facial first impressions.

Chapter 4. Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated

Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (in press). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology*.

4.1. Abstract

The facial first impressions literature has focused on trait dimensions, with less research on how social categories (like gender) may influence first impressions of faces. Yet, social psychological studies have shown the importance of categories like gender in the evaluation of behaviour. We investigated whether face gender affects the positive or negative evaluation of faces in terms of first impressions. In Study 1, we manipulated facial gender stereotypicality, and in Study 2, facial trustworthiness or dominance, and examined the valence of resulting spontaneous descriptions of male and female faces. For both male and female participants, counter-stereotypical (masculine or dominant looking) female faces were perceived more negatively than facially stereotypical male or female faces. In Study 3, we examined how facial dominance and trustworthiness affected rated valence across 1,000 male and female ambient face images, and replicated the finding that dominance is more negatively evaluated for female faces. In Study 4, the same effect was found with short stimulus presentations. These findings integrate the facial first impressions literature with evaluative differences based on social categories.

Keywords: “gender stereotyping” “face perception” “first impressions” “social cognition” “person perception”

4.2. Introduction

4.2.1. First impressions from faces

A wealth of literature has documented that we rapidly and accurately determine social categories from faces, including age, sex and ethnicity (Bruce & Young, 2012). The physical cues underlying such pervasive categories are now relatively well understood (Bruce & Young, 2012), but perceivers also rapidly and consistently infer more abstract social traits from faces such as trustworthiness and dominance (Oosterhof & Todorov, 2008; Willis & Todorov, 2006). These facial first impressions have been shown to affect decisions as important as voting choices (Todorov et al., 2005) and court judgements (Zebrowitz & McDonald, 1991), but less is known about how they are formed.

Recently, however, a substantial advance in facial first impressions research has been to ask participants to rate faces on a range of social traits and then factor-analyse these ratings to uncover underlying dimensions (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009). This approach has revealed three key dimensions of facial first impressions: trustworthiness, dominance, and attractiveness (Oosterhof & Todorov, 2008; Sutherland et al., 2013). The first two dimensions seem to relate to the appraisal of threat, while attractiveness is important in sexual selection. Interestingly, the trustworthiness and dominance facial dimensions may have parallels to warmth (cf. trustworthiness) and competence (cf. dominance) dimensions which underlie the stereotyping of category groups (e.g. Cuddy, Fiske, & Glick, 2008), as well as communal and agentic dimensions found in judgements of the self and others (Abele & Wojciszke, 2007), suggesting the potential for integration between studies of facial first impressions and wider issues in social psychology.

While the dimensional approach has been very useful, the facial first impressions literature has not yet fully investigated the effect of social categories (Quinn & Macrae, 2011). In particular, this facial literature (e.g. Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009) has not looked systematically at how perceptions of these traits or dimensions might be affected by the stereotyping or evaluation of category groups. For example, there are currently no studies of how gender stereotypes affect perceptions of trustworthiness or dominance from faces. Yet, a long history of

social psychological research has shown the importance of social categories and associated stereotypes when evaluating social attributes and behaviour, with gender being an especially salient example (Allport, 1954; Fiske, 2012; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012).

By 'stereotypes', we mean qualities perceived to be associated with particular groups or categories of people (Schneider, 2004). We and others have shown previously that facial representations of groups are stereotypes in the sense that perceivers associate certain facial characteristics with certain groups, and these visual characteristics are closely linked to linguistic stereotypes of those groups (Imhoff, Woelki, Hanke, & Dotsch, 2013; Oldmeadow et al., 2013). This process is similar to the mechanism of "categorization" described by Secord (e.g. Secord, 1958). In categorization, facial first impressions are generated by first assigning a face to a category, and then using associated category knowledge (i.e. stereotypes) to evaluate that face.

This social psychological literature on gender stereotyping, along with the similarities between the trustworthiness/warmth and dominance/competence dimensions found in the social and face perception literatures, suggests that these two dimensions of trustworthiness and dominance may also be evaluated differently in male and female faces, since traits along the first communal dimension are particularly important for females, and traits along the second, agentic dimension are particularly important for males (Cuddy et al., 2008; Prentice & Carranza, 2002; Wiggins, 1979; Wojciszke, 1994). In particular, from social psychological studies showing negative evaluations of targets with counterstereotypical roles or behaviour (e.g. Flannigan, Miles, Quadflieg, & Macrae, 2013; Heilman, 2001; Rudman, 1998), we might expect negative evaluations of targets with counter-stereotypical facial traits.

Instead, the facial first impressions literature links the first, trustworthiness dimension to overall valence, and the second, dominance dimension to perceived femininity-masculinity (e.g. Oosterhof & Todorov, 2008), rather than linking traits on the first dimension to evaluations of femininity and traits on the second dimension to evaluations of masculinity, as in the social literature (e.g. Abele & Wojciszke, 2007; Bem, 1974; Prentice & Carranza, 2002). By hypothesising this separation between valence and facial cues involved in gender stereotypicality (femininity-masculinity),

these facial first impressions models have not yet examined how cues to gender stereotypicality might interact with the gender of the face to influence the valence of facial impressions.

4.2.2. Overview of current studies

An interesting question, therefore, is whether first impressions of faces are also subject to gender stereotyping evaluations. This question brings together social psychological perspectives on gender stereotyping and group-based evaluation with the dimensional facial trait perception approach. Since the social psychological literature has shown that target gender stereotypicality has an impact on perceiver evaluations, here we manipulated facial gender stereotypicality along with facial gender and examined the valence of resulting impressions. We predicted that different sub-groups of faces that differ in gender but which are positioned in the same locations in a dimensional trait face space would be evaluated differently depending on their gender stereotypicality. Specifically, based on previous findings of negative evaluation of counter-stereotypical targets, we investigated whether facially counter-stereotypical women and men would be perceived less positively than facially stereotypical men and women.

To investigate this, we used converging methods. For Study 1 and Study 2, we developed an approach based only on asking participants to write down their thoughts about different faces. For Study 3 and Study 4 we used a complementary but more direct approach of asking participants to rate the valence of their first impressions.

4.3. Study 1

For our first approach to the issue of how the gender-stereotypicality of the face might interact with the valence of facial impressions, we used everyday photographs of real male and female faces with typically feminine or masculine appearance. We investigated spontaneous impressions to these faces by asking participants simply to state whatever came to mind when viewing each face. To visualise these spontaneous impressions we created word clouds depicting the frequency of attributions, and complemented this qualitative approach with quantitative assessments based on coding the participants' descriptions on their valence (positive-negative). The key features of

Study 1 are therefore that the stimuli were real faces that were selected in terms of their gender stereotypicality (feminine or masculine appearance) and that participants' responses were unconstrained. The advantage of this method is that it does not in any way influence what attributions participants make, and therefore shows what is spontaneously perceived.

4.3.1. Methods 1

Ten participants (mean age, 24.3 years; 5 male) took part. Participants provided informed consent to procedures that were approved by the ethics committee of the University of York psychology department. All participants self-identified as culturally Western.

4.3.1.1. Stimuli 1

Twenty faces were chosen from a set of 1,000 highly varied "ambient image" face photographs used in previous studies (Santos & Young, 2005, 2008; Sutherland et al., 2013). The concept of ambient images was introduced by Burton, Jenkins and Schweinberger (2011; see also Jenkins, White, Van Montfort, & Burton, 2011) to emphasise the potential importance of the variability between images of faces we see in everyday life. In order to represent this variability and thus allow us to examine natural first impressions, the ambient image database consists of 1,000 photographs of 500 male and 500 female faces of all adult ages which are deliberately allowed to vary on many cues including pose, expression, lighting and facial paraphernalia such as hairstyles and glasses (Santos & Young, 2005, 2008; Sutherland et al., 2013). The faces were restricted to adults of Caucasian appearance, as cross-cultural or own-race biases were not the focus of this investigation.

From this database of 1,000 ambient images we chose twenty faces for use in Study 1, selecting these on the basis of previously collected sets of ratings of each of the 1,000 images (Santos & Young, 2005, 2008; Sutherland et al., 2013). The twenty faces comprised five men and five women of gender stereotypical appearance (masculine-looking men and feminine-looking women), and five men and five women of gender counter-stereotypical appearance (feminine-looking men and masculine-looking women). Ratings of perceived femininity-masculinity (at least 6 raters per trait, taken

from Santos & Young, 2005, 2008; Sutherland et al., 2013) are presented in **table 4.1.**, together with ratings of other important characteristics. **Table 4.1.** also gives separate ratings of perceived facial femininity (1 low - 7 high) for the female faces and perceived facial masculinity (1 low – 7 high) for the male faces. Note that the faces high and low in gender stereotypicality were deliberately allowed to vary on other traits, so masculine-looking faces of both genders were also rated significantly higher on stereotypically masculine traits (e.g. dominance) and significantly lower on stereotypically feminine traits (e.g. approachability) than feminine-looking faces of both genders. Importantly, the faces were selected to minimise overall trait differences between the genders, so that the male and female face sets did not differ on key gender-stereotypic traits (trustworthiness, approachability, intelligence, smiling and dominance: **see table 4.1.).** This sets up a strong test of whether the gender stereotypicality of the face changes the valence of spontaneous impressions.

Table 4.1. Mean trait ratings (and standard deviations) for the four face sets, at the face level.

	Feminine-looking Females	Masculine-looking Females	Feminine-looking Males	Masculine-looking Males
Femininity-masculinity	1.68 [a c] (0.13)	4.02 [b c] (0.26)	4.20 [a d] (0.53)	6.36 [b d] (0.19)
Femininity (female faces)	5.65 [c] (0.31)	3.44 [c] (1.22)	-	-
Masculinity (male faces)	-	-	2.63 [d] (0.63)	5.58 [d] (0.46)
Approachability	4.93 [c] (1.07)	2.75 [c] (0.87)	4.87 [d] (0.53)	3.38 [d] (1.23)
Trustworthiness	4.80 [c] (0.72)	3.39 [c] (0.68)	4.20 [‡] (0.27)	3.48 [‡] (0.71)
Smiling	4.63 [c] (1.48)	2.27 [c] (1.24)	4.90 (1.58)	2.97 (1.63)
Aggressiveness	2.20 [c] (0.54)	3.82 [c] (1.11)	2.46 (0.54)	3.74 (1.51)
Dominance	3.04 [c] (0.13)	4.48 [c] (0.30)	3.34 [d] (0.34)	4.96 [d] (0.48)
Intelligence	3.92 (0.64)	3.63 (0.94)	3.90 (0.82)	4.06 (0.75)
Attractiveness	5.47 [a c] (0.40)	3.33 [c] (1.04)	4.15 [a d] (0.30)	2.98 [d] (0.85)

[a] significant difference across feminine males and females $p < .05$; [b] significant difference between masculine males and females, $p < .05$; [c] significant difference between female means, $p < .05$; [d] significant difference between male means $p < .05$; [‡] marginally significant difference $p < .07$.

4.3.1.2. Procedure 1

Participants were tested in a quiet room on a PC running E-Prime software (version 2; Psychology Software Tools, Pittsburgh, USA). Faces were shown one at a time with a blank text box. Participants were told that the study was examining first impressions and were invited to type in anything that came to mind on viewing the face, no matter how silly, judgemental or socially inappropriate. They were encouraged to be honest and reassured that their responses were anonymous and that there was no right or wrong answer. Participants were instructed to write as much as possible until they felt that they were no longer being spontaneous. Participants saw the 20 experimental faces in a random order, with two faces - which were not analysed - as a practice.

4.3.1.3. Thematic analysis

Our data were participants' descriptions, split into units consisting of single words or phrases (for example, "not friendly"). Word clouds (wordle.com) were then used to examine themes to see if negative evaluations occurred for counter-stereotypical faces **(see figure 4.1.**). These word clouds present higher frequency descriptions in larger font. This allows exploration of the main concepts emerging from the descriptions. Only key words and short phrases (<5 words) were included, to allow examination of the main concepts without common but uninformative words or long phrases hampering this. Words or phrases were not included where to shorten them or take them out of context would be misleading (e.g. *'she works in a male-dominated area'* could not be fairly summarised).

4.3.1.4. Content analysis

A quantitative content analysis was carried out on all description units to complement and substantiate the qualitative observations. Since the valence of spontaneous impressions was the principal focus of interest, each individual word or phrase was blind coded by two judges for their valence. This was achieved by coding descriptions as either positive or negative if they referred to a skill or problem (e.g. 'sporty', or 'has a difficulty changing her opinions'), used a positive or negative qualifier (e.g. 'good sense of humour' or 'hate his moustache') or were conventionally positive or negative (e.g. 'happy' or 'angry').

4.3.2. Results

The participants produced a total of 2,981 words describing the faces, which were then separated into 1027 units (words and phrases). On average, each person described each face with 5.14 units, suggesting that facial first impressions are relatively rich.

4.3.2.1. Thematic word clouds 1

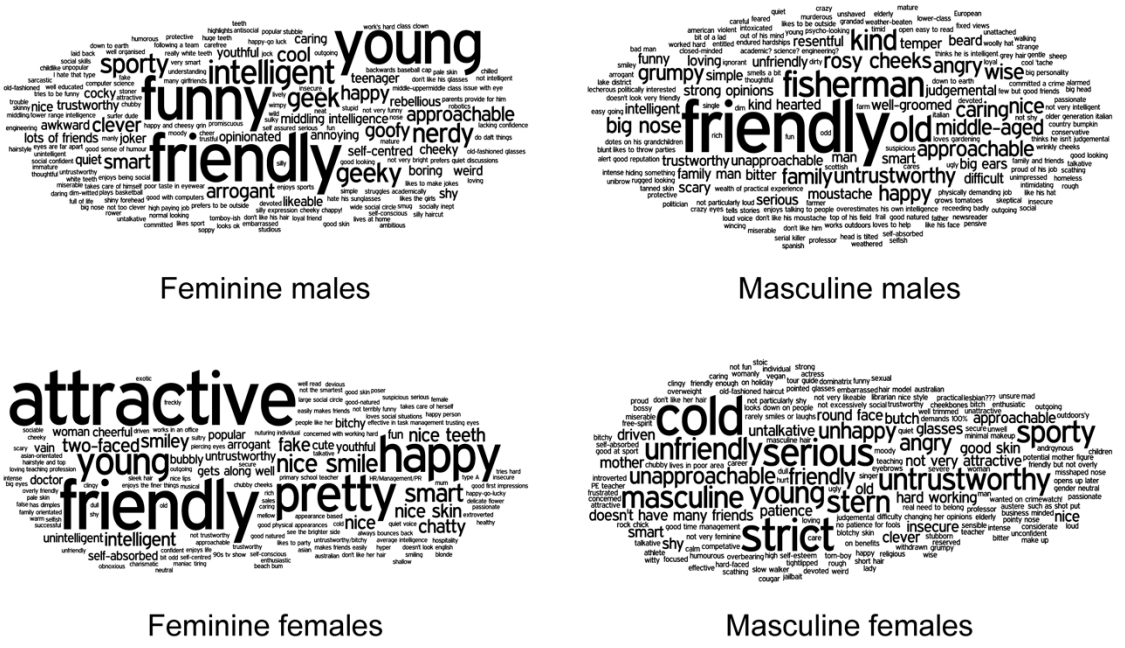

Figure 4.1. Word clouds depicting first impressions made to gender stereotypical (feminine female and masculine male) and gender counter-stereotypical (masculine female and feminine male) faces. Larger font size represents more frequent descriptions.

In total, 109 units (11%) could not be included in the word clouds, either because they lacked a key word or phrase, or because shortening them would change their meaning. The remaining majority (918 units) were inspected for themes for each face group **(see figure 4.1.)**.

The themes emerging for feminine women were largely positive inferences of physical attractiveness, happiness, and friendliness. Masculine men were also described as friendly and approachable, with positive morality impressions (e.g. caring and kind). However, masculine women received mainly negative attributions: strict, cold,

unfriendly and untrustworthy. This negativity was not as evident for feminine males, with themes describing them as funny, friendly and intelligent geeks.

4.3.2.2. Quantitative content analysis

The descriptions were blind coded by two judges for their valence to quantitatively examine the valence of attributions to the different sets of faces. Agreement was high (Kappa = .94, p <.001) and all disagreements were resolved before analysis. We used an index of overall valence by dividing the number of positive words (summed across the 5 faces in the relevant category) by the number of negative words at the participant level. To avoid zero values, the constant 0.5 was added to all cell counts, thus allowing division without error (Gart & Zweifel, 1967).

The data were highly non-normal because they were ratios (no negative values, with a strong rightward skew). Log-transformation of the data still left problems of kurtosis in the residuals (outliers, and high data clustering). We therefore followed Liermann, Steel, Rosing and Guttorp (2004) in their recommendation to use non-parametric tests. Friedman's ANOVA was used to ascertain that there was a reliable difference across the four face groups, followed with planned comparisons using Wilcoxon signed-rank *t*-tests. **Figure 4.2** displays the overall mean ranks because this is what the Friedman's ANOVA is based on, and because the strong skew meant that the median values were also potentially misleading.

The positivity of descriptions was significantly different across the four face categories: X^2 (3) = 19.18, p < .001 **(see figure 4.2).** There was evidence for negative evaluations to counter-stereotypical females: the relatively masculine-looking females received significantly less positive descriptions than the masculine-looking males (Z = -2.40, p < .05) and feminine-looking females (Z = -2.80, p < .005). There was no evidence for negative evaluations to counter-stereotypical males: the relatively masculine and feminine-looking males did not significantly differ (Z = -1.13, p = .30) nor did the more feminine-looking males and females (Z = 1.78, p = .084).

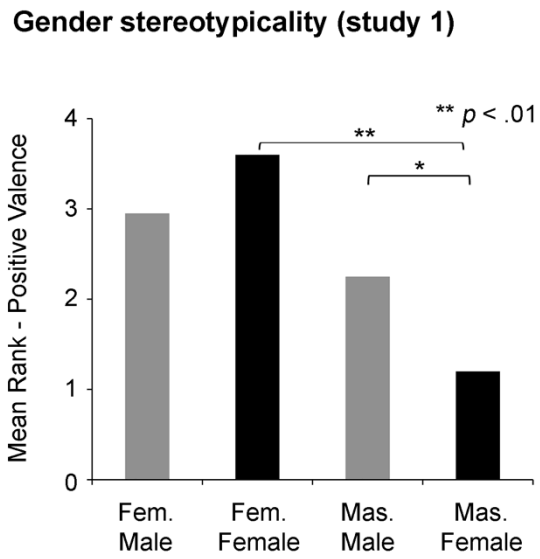

Figure 4.2. Mean ranks from Friedman ANOVAs used to test the overall valence differences for the gender stereotypical and counter-stereotypical faces in Study 1. ** $p < .01$, * $p < .05$.

4.3.3. Discussion 1

Study 1 found evidence that faces were evaluated differently depending on their gender and gender stereotypicality. Each face category elicited distinctive themes, with all groups apart from the counter-stereotypical masculine-looking females receiving largely positive descriptions **(see figure 4.1.)**. As predicted, relatively masculine-looking females received more negative evaluations than the stereotypical masculine males and feminine females. However, there was no evidence for negative evaluation towards the counter-stereotypical feminine-looking males. In terms of models of facial first impressions, this points to the idea that the femininity-masculinity cues theorised to underlie the dominance dimension, can interact with the valence of resulting first impressions, which is theorised to be mainly linked to trustworthiness (cf. Oosterhof and Todorov, 2008).

We should point out that to some extent, these evaluations followed the pre-ratings given to the faces initially **(see table 4.1.)**. Crucially, however, the participants in Study 1 were free to describe anything they liked about the face. That these free descriptions to some extent replicated the previous ratings suggests these impressions really are

salient in a natural sense; and Study 1 is therefore a naturalistic validation of the central point that male and female faces are evaluated differently. Moreover, the male and female faces at a given level of masculinity or femininity did not differ on multiple trait ratings; yet they were evaluated differently by participants.

Clearly, the feminine and masculine faces could (and did) differ on many perceived traits, raising the question of the potential roles of different traits in their evaluation. In particular, since prominent models of facial first impressions focus on trustworthiness and dominance dimensions (Oosterhof & Todorov, 2008), in Studies 2-4, we opted to explore the effect of manipulating facial dominance and trustworthiness on the evaluation of male and female faces.

4.4. Study 2

In Study 2, we examined how male and female faces varying along the two main trait dimensions identified in the facial first impressions literature (trustworthiness and dominance: Oosterhof & Todorov, 2008) were spontaneously evaluated. In order to systematically manipulate these factors and to ensure that descriptions were based only on facial cues underlying the relevant dimension, we used computer-generated images to manipulate perceived facial dominance and trustworthiness separately. Average face stimuli were created by averaging sets of male or female faces from the ambient image database rated high or low on either trustworthiness or dominance. The effect of averaging is to create face-like prototype images (referred to as 'faces' for simplicity) that only include those facial cues that are consistent across the individual faces used to create each average (Mattavelli, Andrews, Asghar, Towler, & Young, 2012; Oldmeadow et al., 2013; Sutherland et al., 2013).

In Study 1, we did not have enough participants to ascertain whether facial evaluations might be affected by gender-specific perceiver biases. A secondary aim of Study 2 was therefore to examine whether participant sex affected the results.

4.4.1. Methods 2

4.4.1.1. Procedure 2

Forty participants (mean age 20.95 years; 20 male) took part. Participants provided informed consent to procedures that were approved by the ethics committee of the University of York psychology department. The experiment was hosted online by Qualtrics (www.qualtrics.com) although participants were recruited through the University. All participants self-identified as culturally Western. Participants saw all eight average face stimuli in a random order and were asked to freely describe them (instructions as before).

4.4.1.2. Stimuli 2

The stimuli were eight carefully controlled computer-generated face-like stimuli (referred to as 'faces' for convenience), created to be high and low on perceived trustworthiness or dominance for each gender **(see figure 4.3.)**. These were taken from pre-existing trustworthiness and dominance stimuli sets consisting of 100 average face images that varied on either trustworthiness or dominance against perceived gender. See Mattavelli et al. (2012) for a full description of the trustworthiness matrix from which the trustworthiness stimuli used in the current study were selected. A similar procedure to that used in Mattavelli et al. (2012) was used to create the dominance matrix used to select the dominance stimuli used in the current study.

Briefly, the 15 most and 15 least dominant male and female faces were selected from the ambient image database, subject to constraints that the photographs included no spectacles, were as close to frontal view as possible, showed no beards or moustaches, and that there were no more than two faces with hats in each set (in order to generate clear stimuli). The faces in each set of 15 photographs were then averaged using PsychoMorph software (Tiddeman et al., 2001) to create four prototype faces. Two quasi-linear continua of 10 male or female face-like images of varying dominance were created by caricaturing or anti-caricaturing each prototype at two levels to increase or decrease its distance from the opposite prototype. These continua of 10 images were then presented in random order and rated for dominance (1–7 scale, low–high) by 14 pilot raters (7 female, mean age 20 years) who did not otherwise participate in the

current study. It was necessary to match continua so that the male and female prototype images were equivalently high or low in dominance. We therefore selected male and female images that were rated equally high or low in perceived dominance. The rest of the 10x10 stimuli matrix was generated by morphing the faces between the prototypes along the dominance and the gender dimensions and adding a caricatured image in each of the four directions, in steps of 14% (see Mattavelli et al. 2012, for a description of the trustworthiness stimuli).

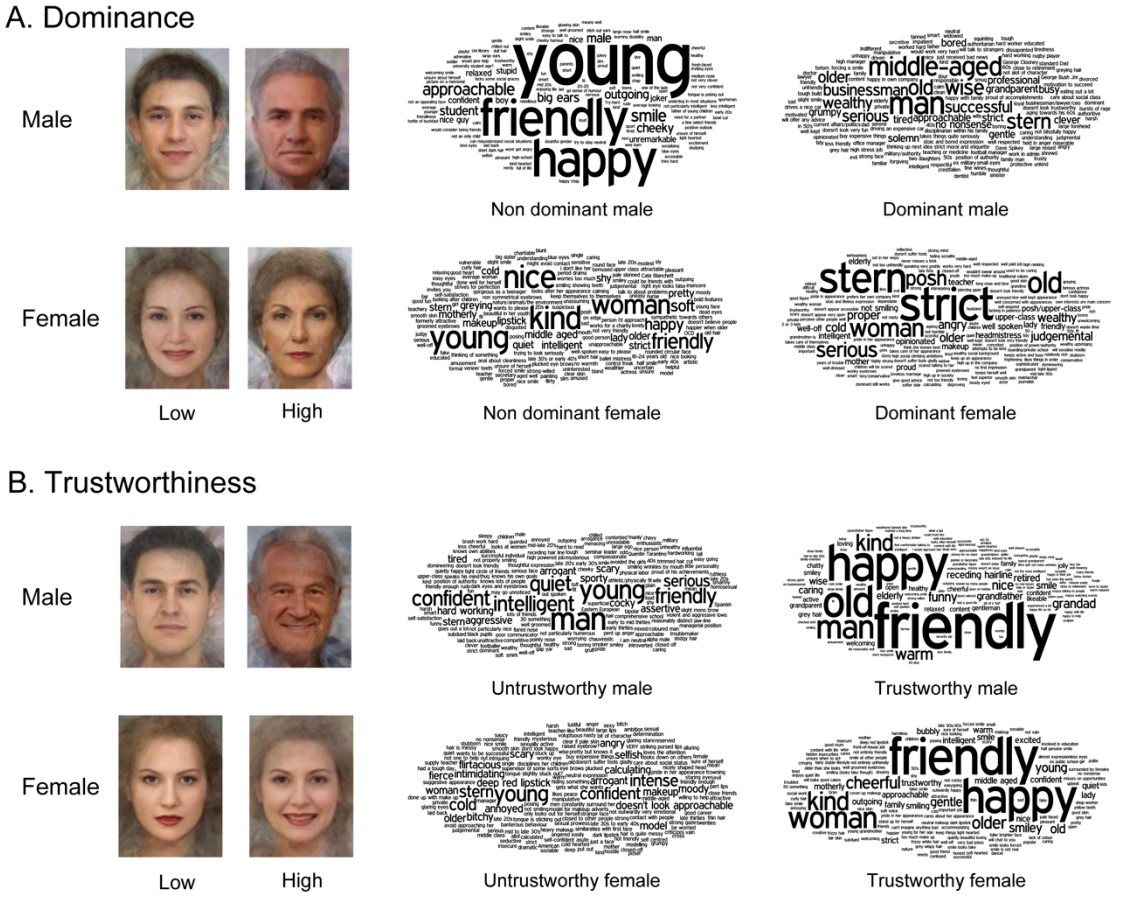

Figure 4.3. Computer-manipulated high and low Dominance (A) and Trustworthiness (B) male and female face stimuli used in Study 2, with word clouds presenting participants' descriptions. Gender stereotypical faces include non-dominant and trustworthy female and dominant and untrustworthy male average faces. Gender counter-stereotypical faces include dominant and trustworthy female and non-dominant and trustworthy male average faces. Larger word cloud font size represents more frequent descriptions.

From these stimuli sets, we had unambiguously male and female images rated on their trustworthiness and dominance by new pilot raters who otherwise did not take part in the current study in order to choose the stimuli used in the current study (n = 10, 5 female; mean age: 25 years). Crucially, we chose eight faces (**see figure 4.3.,** male and female, high and low on perceived dominance or trustworthiness) so that high and low dominance and trustworthiness prototype faces did not differ in rated traits across gender, and so that the dominance faces varied more on dominance than the trustworthiness set and vice versa **(see table 4.2.)**.

Table 4.2. Mean trait ratings (and standard deviations) for the face-like images used as stimuli in Study 2 (shown in figure 4.3.), at the participant level.

Face set	Face characteristics		Trust.	Dom.
Dominance	Low	female	5.1 (1.45)	2.4 (0.70) [a]
		male	5.2 (0.92)	2.0 (0.82) [b]
	High	female	4.2 (1.69)	5.4 (1.10) [a]
		male	4.0 (1.41)	5.8 (0.63) [b]
Trustworthiness	Low	female	3.5 (0.97) [c]	4.8 (1.40) [c]
		male	3.5 (0.85) [d]	4.4 (0.97)
	High	female	5.2 (1.32) [c]	2.9 (1.20) [c]
		male	5.6 (1.43) [d]	4.1 (2.08)

[a] significant difference between low and high dominance females $p < .05$; [b] significant difference between low and high dominance males, $p < .05$; [c] significant difference between low and high trustworthy females $p < .05$; [d] significant difference between low and high trustworthy males $p < .05$. All other differences $p > .07$.

4.4.2. Results 2

Overall, participants used 6,843 words to describe the faces. Descriptions could be divided into 1,778 units (words and phrases; 5.56 units on average per participant and face).

All data procedures were as for Study 1. As expected, the data (and residuals) were highly non-normal due to a strong positive skew, and log-transformation was not completely satisfactory due to high clustering and outliers, so we followed Liermann et al. (2004) and used non-parametric tests. **Figure 4.4** again plots the mean ranks from the Friedman ANOVAs. A second individual coded 75% of the data and agreement was

high (Kappa = .74, p <.001). All disagreements were resolved before analysis. Descriptions given to trustworthiness and dominance faces were analysed separately.

4.4.2.1. Thematic word clouds 2: Dominance faces

In total, 73 units (8%) could not be included. The majority (809 units) were entered into word clouds and examined for themes. These echoed Study 1 **(see figure 4.3.)**. Specifically, the counter-stereotypical (dominant) female face primarily received negative themes of strictness or sternness. In contrast, the dominant male face was described as a patriarchal, largely benevolent leader (e.g. as a wise businessman). The non-dominant female received stereotypically feminine descriptions, focusing on her as kind, nice, and motherly. Finally, the non-dominant male face was seen as friendly and outgoing.

4.4.2.2. Content analysis 2: Dominance faces

We used Mann-Whitney U tests to ascertain whether there was a difference in the valence of the descriptions made by female and male participants. No significant differences were found (all $U > 145$, $p > .1$) so the data were pooled across participants **(see figure 4.4).**

The overall ANOVA was significant: X^2 (3) = 20.01, $p < .001$. We found that the dominant (counter-stereotypical) female face was described less positively than the dominant male (Z = -2.76, $p < .01$) and non-dominant female face (Z = -3.18, $p < .005$). As in Study 1, there was no evidence for negative evaluations of the counter-stereotypical male, with no significant differences in valence between the non-dominant male and female faces (Z = -0.60, p = .56) or the dominant and non-dominant male faces (Z = -1.45, p = .15). In sum, the findings for the dominant faces were highly similar to Study 1.

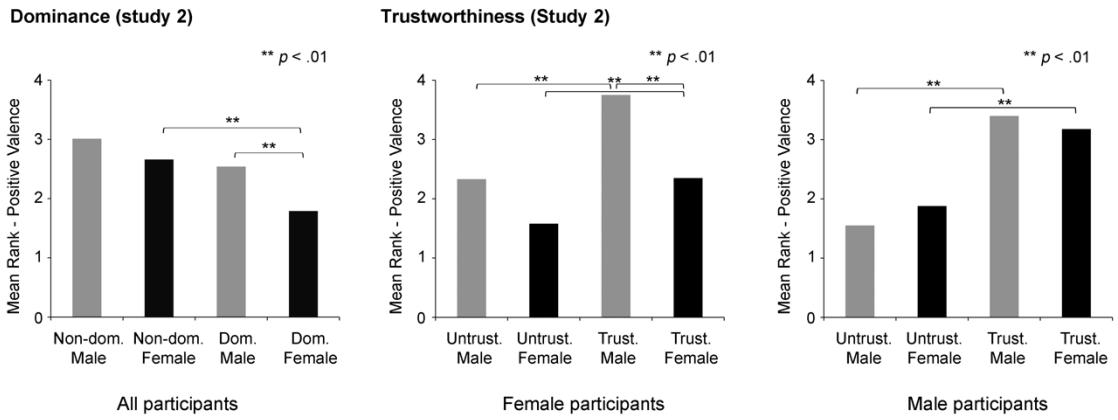

Figure 4.4. Mean ranks from Friedman ANOVAs used to test the overall valence differences for the male and female dominance and trustworthiness faces in Study 2. Male and female participants' responses differed to trustworthy faces in Study 2, so are presented separately. ** $p < .01$, * $p < .05$.

4.4.2.3. Thematic word clouds 2: Trustworthiness faces

In total, 75 units (8%) could not be included; the majority (821 units) were entered into word clouds. Themes mainly reflected a large difference between trustworthy and untrustworthy faces, supporting Oosterhof and Todorov's (2008) suggestion that valence is highly linked to perceived trustworthiness **(see figure 4.3.)**. The trustworthy male received highly positive descriptions centring on his friendliness, kindness and age. The trustworthy female face also received similar positive descriptions, although some participants thought that her friendliness was socially polite rather than genuine. The untrustworthy male was described somewhat ambivalently; although some participants described him as friendly and intelligent, others described him as arrogant, stern and scary. Finally, the untrustworthy female was described mainly negatively, as cold and intimidating.

4.4.2.4. Content analysis 2: Trustworthiness faces

In general, male and female participants gave similar descriptions: the majority of Mann-Whitney *U* tests comparing participant gender were not significant ($U > 138$, $p > .09$). However, female participants described the trustworthy and untrustworthy male faces in significantly more positive terms than the male participants. The female participant mean rank for the trustworthy male face was 24.45, with the male participant

mean rank as 16.55 ($U = 121$, $p < .05$). The female participant mean rank for the untrustworthy male face was 24.53 compared to the male participant mean rank of 16.48 ($U = 119.50$, $p < .05$). Consequently, we analysed the trustworthiness descriptions separately by participant gender **(see figure 4.4).**

The overall ANOVAs were significant for both male (X^2 (3) = 31.98, $p < .001$) and female participants (X^2 (3) = 31.22, $p < .001$). For male participants, the untrustworthy faces were described in significantly less positive terms than the trustworthy faces (male faces: $Z = -3.92$, $p < .001$; female faces: $Z = -3.06$, $p < .005$). Similarly, for female participants the untrustworthy faces were described in significantly less positive terms than the trustworthy faces (male faces: $Z = -3.29$, $p < .001$; female faces: $Z = -2.55$, $p < .01$). There was no evidence of differentially negative evaluations directed at the counter-stereotypical female face, as the untrustworthy female and male faces did not significantly differ for either female ($Z = -1.78$, $p = .078$) or male participants ($Z = -0.63$, $p = .54$). Interestingly, the female participants also gave significantly more positive descriptions to the high trustworthy male face than the high trustworthy female face ($Z = 3.36$, $p < .001$). This comparison was not significant for the male participants ($Z = 0.51$, $p = .63$).

4.4.3. Discussion 2

In Study 2, it was again clear that different subgroups of faces defined by their gender and perceived social traits were evaluated in different ways. The faces that differed in dominance, but not the faces that differed in trustworthiness, closely followed the pattern of results in Study 1. That is, the counter-stereotypical (dominant-looking) female face was described in less positive terms than the stereotypical (non-dominant) female face and the stereotypical (dominant) male face. Again, there was no evidence for differentially negative evaluations directed at the counter-stereotypical male face.

For the faces manipulated on trustworthiness, the counter-stereotypical (trustworthy) male face was actually described in more positive terms than the stereotypical (trustworthy) female face, at least by female participants. Interestingly, this is contrary to the "women are wonderful effect" described in social psychology, where women are generally perceived more positively than men because they are perceived as higher in

stereotypically feminine, pro-social traits (Eagly & Mladinic, 1994). Nevertheless, both the female and male faces high on perceived trustworthiness received more positive evaluations than those low on perceived trustworthiness, supporting Oosterhof and Todorov's (2008) suggestion that the valence of first impressions is strongly linked to perceived trustworthiness.

Using computer-manipulated average faces in Study 2 offers a well-controlled test because the image manipulation techniques remove any variance between the face photographs that does not contribute to the trait of interest (e.g. extra-facial paraphernalia), while allowing consistent cues to remain. This minimises the possibility that extra-facial or facially idiosyncratic characteristics caused the findings. Of course, the face averages differ on a variety of cues. For example, inspection of **figure 4.3.** shows that age and expression contribute to some extent to perceived dominance and trustworthiness. This is expected because Zebrowitz and colleagues have shown that facial maturity is linked to dominance (e.g. Berry & McArthur, 1986) and Todorov and colleagues have shown that facial expression is closely linked to trustworthiness (e.g. Todorov, Baron, & Oosterhof, 2008). Here we took the approach of allowing all of the facial cues to trustworthiness or dominance to naturally contribute to these trait judgements, rather than attempting to 'control' these cues. Indeed, controlling these cues would change the judgement being made, and we wanted to examine trustworthiness and dominance judgements as they are naturally made. We note that discussing these results in terms of trait (e.g. trustworthiness) judgements or in terms of cues to trait judgements (e.g. happiness expressions) does not change the interpretation of the results (see Keefe, Dzhelyova, Perrett, & Barraclough, 2013). Instead, these facial cues are intrinsic to the trait judgements being made. The important points are that the averaged faces differed systematically in the perceived traits of trustworthiness and dominance and were matched as closely as possible across face gender.

4.5. Study 3

In the previous two studies, we deliberately used a small sample of faces in order not to overburden participants with the free descriptions task and/or because they were prototypical average images. In addition, the method used in Study 1 and Study 2

involved independent raters coding the valence of participants' spontaneous descriptions rather than asking participants to note the valence of their own impressions. The advantage of using spontaneous descriptions is that participants are not offered any suggestions as to what might be important evaluative dimensions. However, it would be useful to ascertain that our results held with a larger sample of faces and a direct judgment of the valence of participants' first impressions.

Participants in Study 3 therefore rated the full ambient image database set (1,000 faces) on the valence of their first impression (i.e. how negative or positive their first impression was). We then used regression to predict these valence ratings from ratings of dominance or trustworthiness previously collected from an independent sample of raters (see Sutherland et al., 2013). From the pattern of results across Studies 1 and 2, we expected to find that the valence of the impression depended on an interaction between the dominance of the face and the gender of the face, so that dominance would be evaluated more negatively for female faces compared to male faces. However, we expected to find that the trustworthiness of the face was strongly related to the valence of the first impression for both male and female faces, based on Oosterhof and Todorov's (2008) claim that the first dimension is heavily valenced and also on our finding in Study 2 that both male and female faces high in trustworthiness were perceived more positively than those low in trustworthiness.

4.5.1. Method 3

Twelve participants (mean age 18.67 years; 6 female) took part. Participants provided informed consent to procedures that were approved by the ethics committee of the University of York psychology department. All participants self-identified as culturally Western.

4.5.1.1. Stimuli 3

The participants viewed all 1,000 faces (500 male) from the ambient image database used to construct the stimuli in Studies 1 and 2 (see Methods 1). The female faces in this database (k = 500) are rated as significantly more trustworthy than the male faces (t(981.70) = 13.69, p <.001) with a mean trustworthiness rating of 4.37 (0.66 *SD*) compared to the male faces at 3.75 (0.76 *SD*). The female faces are also rated as

significantly less dominant than the male faces ($t(998) = -35.72$, $p < .001$) with a mean dominance rating of 3.31 (0.61 *SD*) compared to the male faces at 4.73 (0.65 *SD*). The facial ratings of trustworthiness and dominance therefore parallel extra-facial gender stereotypes (e.g. Prentice & Carranza, 2002).

Participants were tested in a quiet room on a PC running E-Prime software (version 2; Psychology Software Tools, Pittsburgh, USA). Faces appeared in random order with a valence rating scale underneath. Participants were instructed to rate the faces on a scale of 1-7 for how positive their first impression of the person was, with 1 being labelled as a very negative impression, and 7 as a very positive impression. They pressed the number key that corresponded with their rating and the next face photograph then appeared after a blank interval of approximately 750ms. Face photographs were 150 pixels in height (approximately 5cm on screen) with varying width to preserve aspect ratio. Given that 1,000 faces had to be rated, participants were allowed to take a break whenever they wished. They were given as much time as they wanted to look at each face, but were encouraged to go with their 'gut instinct' (Todorov et al., 2005). They first saw 10 faces, randomly drawn from the database, as a practice.

4.5.2. Results 3

The reliability of the valence ratings was good, with Cronbach's alpha = .85 ($p < .001$). These valence ratings were correlated to previously collected ratings of dominance and trustworthiness, for male and female faces separately (Santos & Young, 2005, 2008; Sutherland et al., 2013). At least six raters had previously rated each trait and all trait reliabilities (alphas) were above .7 (see Santos & Young, 2005, 2008; Sutherland et al., 2013 for further details). All statistical comparisons of correlation coefficients used Fishers' *r*-to-*z* transform.

We employed hierarchical regression analyses to assess the interactions between the trait ratings and gender in predicting the valence of first impressions. (i.e. moderated regression with a categorical predictor using dummy coding: Aiken & West, 1991; Warner, 2013). Here, the interaction is calculated as the product of the moderator and independent variable, and directly entered as a term in the regression (Aiken & West, 1991; Warner, 2013). For dominance, we first used the face gender (0,1) and dominance

rating (centred) to predict the valence ratings, in order to test these main effects. **Figure 4.5.** shows the scatterplot of dominance ratings against valence ratings. In a second block we included the interaction of dominance and face gender (0,1) to predict the valence ratings in order to directly test the interaction. The dominance and gender main effects were significant **(see table 4.3.)**, demonstrating that female faces and faces low in dominance were linked to more positive evaluations. Importantly, the interaction between dominance and gender was also significant and explained a significant proportion of the variance over and above the main effects **(see table 4.3.**). To examine this significant interaction, we ran separate regressions predicting the valence of first impressions from dominance for male and female faces separately. For female faces, dominance (centred) significantly negatively predicted the valence of impressions: beta = -.36, $t(498) = -8.50$, $p < .001$, and also explained a significant proportion of the variance of the valence of impressions: $R^2 = .13$, $F(1, 498) = 72.21$, $p < .001$. However, for male faces, dominance (centred) significantly positively predicted the valence of impressions: beta = .12, $t(498) = 2.58$, $p = .01$, and explained a small but statistically significant amount of the variance in the valence of impressions: $R^2 = .01$, $F(1, 498) = 6.65$, $p = .01$. Our prediction that dominance would be more negatively evaluated for female faces compared to male faces was therefore supported (see **figure 4.5.**).

Table 4.3. R^2 and standardised beta weights for individual predictors in predicting valence. Traits were centred. Results are shown for Study 3 (unlimited presentation) and Study 4 (500ms presentation)

		Valence		Valence 500ms	
		Dominance	Trust.	Dominance	Trust.
Step 1	Trait (dominance/ trustworthiness)	-.141*	.864**	-.257**	.852**
	Gender	-.249**	-.011	-.141*	.005
Step 2	Trait*Gender	.397**	.029	.271**	-.037
	R^2 change	.046**	.000	.022**	.001
	Total R^2	.181**	.754**	.162**	.723**

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

We then used the face gender (0,1) and trustworthiness ratings (centred) to predict the valence ratings, in order to test these main effects. **Figure 4.5** shows the scatterplot of trustworthiness ratings against valence ratings. In a second block we included the

interaction of trustworthiness and face gender (0,1) to predict the valence ratings in order to directly to test the interaction. The main effect of trustworthiness was positive and highly significant, but the main effect of face gender and the interaction between trustworthiness and gender were not significant and the interaction did not explain any more variance over and above the trustworthiness rating (**see table 4.3.**). This shows that the valence of first impressions is highly positively linked to the perceived trustworthiness of both male and female faces (see **figure 4.5**), supporting Oosterhof and Todorov (2008).

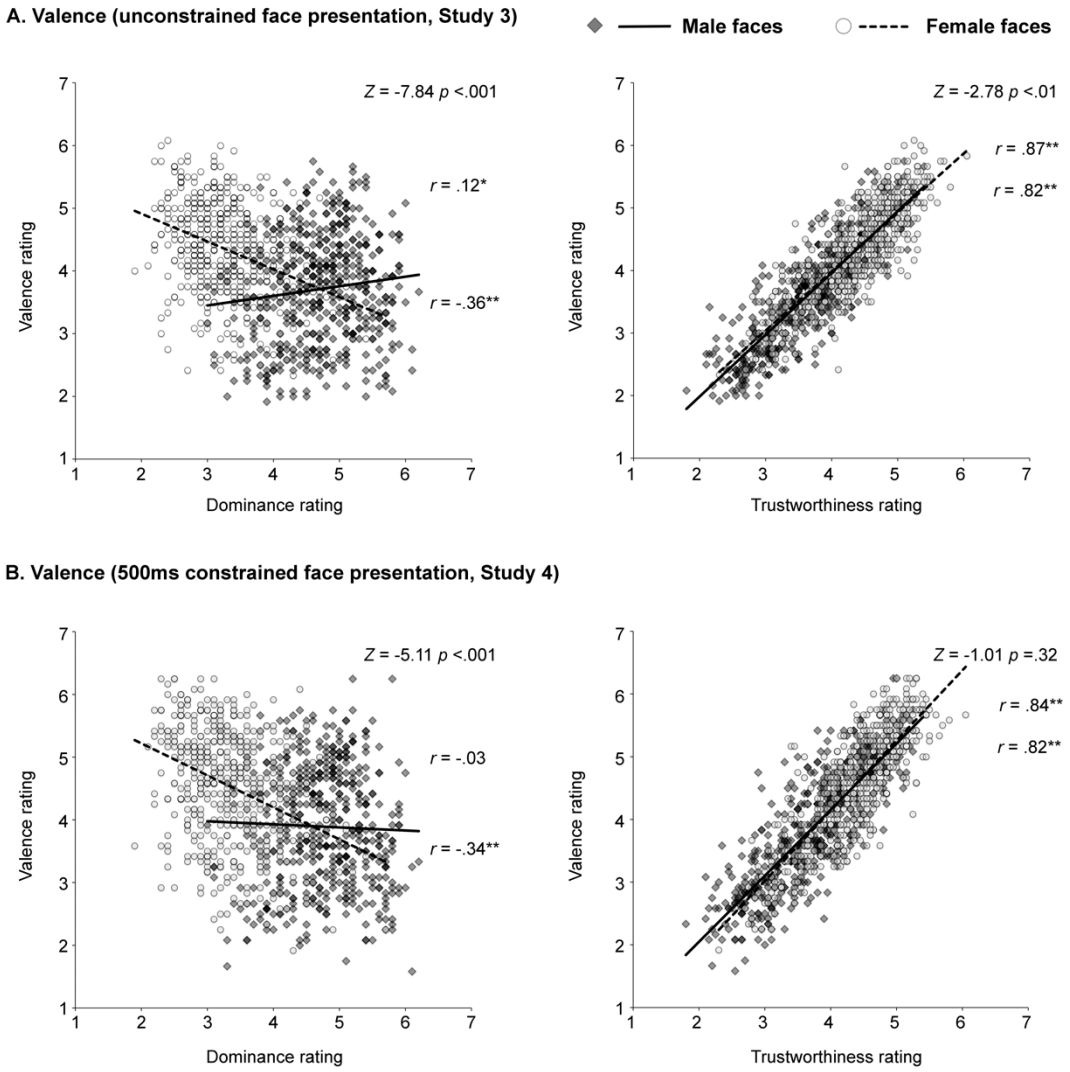

Figure 4.5. Correlations between dominance and trustworthiness ratings for 500 male and 500 female ambient image faces with the valence of impressions made under A) unconstrained viewing conditions (Study 3) and B) 500ms viewing time (Study 4). ** p <.001, * p <.05.

4.5.3. Discussion 3

As predicted, there was an interaction between facial dominance and facial gender which significantly predicted the valence of the first impression, so that dominance was evaluated more negatively for female faces compared to male faces. Dominance was also negatively correlated to the valence of the first impression for female faces, whereas for male faces the correlation between dominance and valence was slightly positive. This supports the findings of Study 2 using a large database of 1,000 highly variable images and a direct measure of perceived valence. We also found that trustworthiness correlated strongly with valence for both male and female ambient image faces, as would be predicted by Oosterhof and Todorov (2008).

4.6. Study 4

In Studies 1-3, we did not control the length of time people could view the faces and respond. This was deliberately done to ensure that we recorded all initial impressions and to keep the task as natural as possible. Moreover, this follows the procedure of many studies in the facial first impressions literature (e.g. Oosterhof & Todorov, 2008; Todorov et al., 2005). However, one could argue that if the differential evaluations we noted in Studies 1-3 truly reflect *first* impressions, then the overall pattern of findings should remain even with short stimulus presentation, since studies have shown that facial first impressions made to faces shown for less than a second largely correspond to unconstrained facial first impressions (Todorov et al., 2009; Willis & Todorov, 2006). Therefore, in Study 4 we repeated Study 3 but restricting the presentation time for each stimulus to 500ms and following this with a disruptive pattern mask. We chose 500ms since this lies in the middle of the range of presentation times used by previous investigations of facial first impressions (Willis & Todorov, 2006). We aimed to replicate the finding that gender and dominance interacted to predict the valence of the first impression to faces seen for less than a second.

4.6.1. Methods 4

Twelve participants (mean age, 19.83 years; 6 male) took part. Participants provided informed consent to procedures that were approved by the ethics committee of the

University of York psychology department. All participants self-identified as culturally Western. The stimuli and procedures were identical to Study 3, except now the faces appeared for only 500ms before being replaced by a visual mask. The masking stimulus was a mosaic scrambled face, constructed by jumbling segments of the average face taken over the whole database, following Todorov et al. (2009). The faces were presented at 150 pixels in height (approximately 5cm) with varying width to preserve aspect ratio and the visual mask was the size of the largest image (150 by 160 pixels).

Participants rated all 1,000 faces on the valence of their first impression (negative-positive). As in Study 3, faces appeared in random order, but trials were divided into five blocks with breaks to rest in between, since the way the Study 4 had to be set up meant that (unlike Study 3) participants could not take a break at any point. Participants were allowed to rate each face after the onset of the visual mask if necessary (following (Todorov et al., 2009) but were encouraged to go with their 'gut instinct' (Todorov et al., 2005) and to try to respond while the face was still visible. All other aspects of presentation were as Study 3.

4.6.2. Results 4

The reliability of the valence ratings at 500ms face presentation was good, with Cronbach's alpha = .88 (p <.001). Moreover, the 500ms valence ratings correlated highly with the unconstrained valence ratings from Study 3: $r_{\text{female faces}}$ = .84 (p <.001), $r_{\text{male faces}}$ = .85 (p <.001). All statistical comparisons of correlation coefficients used Fishers' r-to-z transform.

Again, we used previously collected ratings of dominance and trustworthiness (as for Study 3) to predict the valence ratings (now from a 500ms viewing of the face) for male and female faces separately. **Figure 4.5** shows the scatterplot of dominance and trustworthiness ratings against valence ratings.

To test the interactions between trait impressions and the gender of the face directly, we repeated the hierarchical regression procedure used in Study 3. We first used the perceived facial dominance ratings (centred) and face gender (0,1) to predict the valence of impressions to time-limited faces, and then entered the interaction between

dominance and gender as a predictor in a second block. As in Study 3, the main effects of face gender and dominance were both significant, so that having a female face and looking low in perceived dominance were linked to more positive evaluations **(see table 4.3.)**. Importantly, as in Study 3, the interaction between facial dominance and gender was significant and predicted a significant proportion of the variance in valence impressions over the main effects (**see table 4.3.**). To examine this significant interaction, we ran separate regressions predicting the valence of first impressions from dominance for male and female faces separately. For female faces, dominance (centred) significantly negatively predicted the valence of impressions: beta = -.34, $t(498) = -8.19$, $p < .001$, and explained a significant proportion of the variance of the valence of impressions: $R^2 = .12$, $F(1, 498) = 67.03$, $p < .001$. However, for male faces, dominance (centred) did not significantly predict the valence of impressions: beta = -.03, $t(498) = -0.75$, $p = .454$; or the variance in the valence of impressions: $R^2 = .001$, $F(1, 498) = 0.56$, $p = .454$. Our finding from Study 3 that dominance would be more negatively evaluated for female faces compared to male faces was therefore replicated in Study 4.

We also used the perceived facial trustworthiness ratings (centred) and face gender (0,1) to predict the valence of impressions to time-limited faces, and then entered the interaction between trustworthiness and gender as a predictor in a second block. As in Study 3, the main effect of trustworthiness was positive and highly significant, but the main effect of face gender and their interaction were not significant (**see table 4.3.**). Again, this shows that the valence of first impressions is highly positively linked to the perceived trustworthiness of both male and female faces (see also **figure 4.5.**).

4.6.3. Sex of participant

Since the ratings were highly correlated between Study 3 and 4, we pooled the participants to enable an investigation of participant sex differences. We did this in order to demonstrate that our effects were not driven by one sex (e.g. they were not restricted to male participants' judgements, for example). Here, our question was whether perceived dominance or trustworthiness would affect the valence of first impressions made by male ($n = 12$; alpha = .80) or female participants ($n = 12$; alpha = .92) differently. The main findings remained the same for female and male participants' valence ratings when analysed separately. The interaction between facial dominance

and gender was significant and predicted a significant proportion of the variance in the valence of first impressions over the main effects, so that dominance was evaluated more negatively in female faces for both male and female participants (both $p < .001$). For both male and female participants, dominance (centred) significantly negatively predicted the valence of impressions made to female faces (both $p < .001$) but not to male faces (both $p > .26$). For both male and female participants, as expected, the interaction between trustworthiness and gender was not significant in predicting the valence of first impressions (both $p > .7$).

4.6.4. Discussion 4

In Study 4 we again found that perceived dominance was more negatively valenced for female faces than male faces, even when the valence impression was made with only 500ms exposure to the faces. This confirms that social categories can influence the first impression of facial trait dimensions. Moreover, we also found that trustworthiness was again highly correlated with valence at 500ms exposure for both male and female ambient image faces, as predicted by Oosterhof and Todorov (2008). These results held across the pooled data from Studies 3 and 4 for male and female participants analysed separately. Finally, we found that unconstrained and 500ms constrained ratings of the valence of overall impressions were highly correlated, adding to the literature showing that first impressions can be formed reliably from time-limited presentation of face stimuli (e.g. Willis & Todorov, 2006).

4.7. General discussion

We examined how gender stereotypes affect the valence of first impressions along fundamental dimensions of social face perception. Across four studies, we found that qualitative descriptions and quantitative ratings of faces differed in valence depending on the gender category and perceived social traits of the face. In particular, facially more masculine-looking and more dominant-looking female faces were described and rated more negatively than their male counterparts. Importantly, this pattern of findings was consistent across approaches involving unconstrained responses (Study 1 and Study 2) and a more conventional rating procedure (Study 3 and Study 4). This convergence of findings across such different methods strengthens the claim that it represents an

important phenomenon. It is also worth highlighting that most of our findings were similar across male and female participants. In particular, the negative evaluations of relatively dominant-looking female faces came from women as much as from men.

While the dimensional approach in facial first impressions (e.g. Oosterhof & Todorov, 2008; Sutherland et al. 2013) has made important theoretical advances in understanding the nature of these first impressions (Bruce & Young, 2012), the current research points to the need to integrate social category groups with facial trait dimensions. Certainly, the trustworthiness dimension explained most of the variance in the valence of the first impressions, supporting Oosterhof and Todorov's (2008) claim that the trustworthiness dimension is primarily a judgement of valence. However, we also found clear evaluative differences based on gender category groups along the dominance dimension so that counter-stereotypical female faces were perceived more negatively than stereotypical male or female faces. In contrast, we found little evidence that counter-stereotypical male faces were perceived more negatively than stereotypical male or female faces.

Having demonstrated such clear gender-based evaluative differences, we need to consider potential mechanisms underlying them. Our findings suggest a number of possible avenues for future investigation. One possibility is that these evaluations could represent a simple association between facial cues and gender stereotypes linked to evaluations, similar to the finding that more prototypically 'black' faces are associated with congruent racial stereotypes and negative evaluations (e.g. Livingston & Brewer, 2002). This explanation is similar to the categorization mechanism of facial first impressions described by Secord (1958). In categorization, if perceivers can assign a face to a category, they will then use knowledge associated with that category (stereotypes) to make inferences about the face.

An alternative account relates the evaluations of the faces to potential differences in the frequencies with which they might be encountered. Since women are rated as less facially dominant and more facially trustworthy than men (see Study 3), it is possible that familiarity or perceptual or categorisation fluency might play a role in the negative evaluation of dominant and masculine female faces if these counter-stereotypical women are less frequently encountered or less easily categorised than their stereotypical

counterparts. However, we note that this alternative account would not explain the relatively favourable evaluations of the feminine, non-dominant or trustworthy-looking men, who under this account would presumably also be less familiar or less easily categorised. Moreover, these stereotyping and familiarity/fluency explanations need not be mutually exclusive: counter-stereotypical faces may be less familiar due to their inherent facial structural features, or due to their lesser or biased representation in our cultural lives (e.g. perhaps as result of the under-representation of female leaders in general or in the media: Cracknell, 2013). By calling these 'stereotypes' we are not implying that they are arbitrary.

The specificity of the negative evaluations of dominant women might also index a 'backlash' effect paralleling social psychological findings that women who behave counter-stereotypically (for example, assertively or dominantly) are disliked and receive discrimination in the form of less favourable hiring decisions compared to their stereotypical male and feminine female counterparts (e.g. Heilman, 2001; Rudman & Glick, 2001; Rudman, 1998). However, Sczesny and colleagues have found that, in the context of employment decisions, rather than facing discrimination, facially masculine women are preferred to facially feminine women for a masculine-typed occupation (Sczesny, Spreemann, & Stahlberg, 2006; Sczesny & Kühnen, 2004; Lammers, Gordijn, & Otten, 2009). It may be that facial backlash for dominant women occurs only for overall target evaluation, rather than hiring decisions (and we note that Amanatullah & Tinsley, 2013 found this pattern from behavioural, non-facial target judgements).

If our results represent facial backlash, then it is currently unclear if this is similarly motivated to non-facial backlash (e.g. by system justifying beliefs: Rudman et al., 2012). Moreover, the concept of ambivalent sexism parallels backlash by describing how sexists can be both hostile and benevolent towards women, depending on whether or not they conform to traditional roles (Glick & Fiske, 1996, 1997). It would be interesting in future to examine whether individual differences such as the strength of system-justifying beliefs or the presence of ambivalent sexism moderate the strength of these evaluations, as backlash or ambivalent sexism accounts would predict.

Finally, we focused on global dimensions of facial first impressions across all four studies; gender stereotypicality, trustworthiness and dominance. However, it is worth

noting that a multitude of cues (for example, expression, attractiveness or age) may contribute towards the perception of these traits. In the current studies, we allowed the cues within faces to vary naturally in order to create the perception of main traits of interest (such as dominance), and then examined the consequences of these cues (the evaluation of these faces). We did this in order to let the cues contributing to judgments of interest vary naturally. Indeed, controlling for a cue to a trait necessarily changes the perception of the trait being studied, in turn changing the effect in question. Rather than examine cues or traits individually, we reasoned that the most theoretically powerful approach was to focus on the main dimensions implicated in the facial first impressions literature, and examine these as naturalistically as possible. As Keefe and colleagues have argued (Keefe et al., 2013), one of the consequences of studying complex judgements such as trustworthiness is that one naturally also studies the effects of the cues towards such judgements (such as happiness expressions). Future work may none the less wish to try to isolate the influence of particular cues.

4.8. Conclusions

Using diverse methods to examine the valence of facial first impressions, the current set of studies demonstrate that how traits are evaluated in faces depends on the gender and stereotypicality of the target face. Specifically, we found that women with counter-stereotypical (less feminine or more dominant-looking) faces were evaluated more negatively than men or women with stereotypical faces. For men, being facially counter-stereotypical in trustworthiness or dominance did not lead to such negative evaluations. These findings are theoretically important in bringing together the face perception and social psychological literatures by indicating the need to integrate social category representations with dimensional accounts of facial first impressions.

Chapter 5. Chinese impressions of Asian and Western faces

5.1. Abstract

Studies of Western perceivers have found a three-factor structure to first impressions of Caucasian faces: approachability, youthful-attractiveness and dominance (Sutherland et al., 2013). These factors are thought likely to reflect responses to facial cues with a long evolutionary history related to threat and to sexual selection (Oosterhof & Todorov, 2008; Sutherland et al., 2013). If this is the case, these dimensions should exist in some form across different cultures. However, as yet research has only examined models derived from Western first impressions. To address this, we sought to derive models of first impressions of Asian and Caucasian faces from Chinese perceivers' spontaneous judgments. In Study 1, we asked 20 Chinese participants to freely describe 60 Asian face photographs. Participants spontaneously inferred character traits, especially approachability-related traits. In Study 2, 120 Chinese participants rated a set of 500 Asian and 500 Caucasian face photographs on the twelve characteristics most frequently mentioned in Study 1. We found a three-factor structure for Chinese impressions of Caucasian faces, in which the first two factors were very similar to the first two factors of Western perceivers' impressions of Caucasian faces (approachability and youthful-attractiveness: Sutherland et al., 2013). For Asian faces, we found four factors, including an additional attractiveness factor, indicating that there are also culture-specific aspects of facial evaluation.

Keywords: "face perception" "first impressions" "cross culture"

5.2. Introduction

5.2.1. Facial first impression models

When meeting a stranger for the first time, one of the most salient sources of information we have is their face (Bruce & Young, 2012). Not only does someone's face reveal their gender and approximate age, it can give the perceiver an insight into their ethnicity and attractiveness (Bruce & Young, 2012). However, perceivers go further than these relatively objective judgements, and also frequently infer character traits from facial cues (Todorov et al., 2005). Moreover, there is an association between these facial first impressions and important decisions such as whether to lend money to the target (Duarte et al., 2012; X. Yang, 2014), whether to allow them to win a court case (Zebrowitz & McDonald, 1991) and even whether to elect them to political office (e.g. Antonakis & Dalgas, 2009; Mattes & Milazzo, 2014; Todorov, Mandisodza, Goren, & Hall, 2005). These real-life findings and the considerable public interest in facial first impressions – for example, a Google search for 'facial first impressions' generates over 5 million hits – make this research area highly deserving of further scientific attention.

Recently, researchers have attempted to characterise the key dimensions underlying facial first impressions (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009). In a pioneering study, Oosterhof and Todorov (2008) collected participants' spontaneous descriptions of a set of 66 face photographs, categorised these descriptions into traits, and then had these face photographs rated along with computer generated faces on the most frequently mentioned categories of traits. They used a principal components analysis to reduce these facial trait ratings into key underlying dimensions, finding two such dimensions: trustworthiness and dominance (Oosterhof & Todorov, 2008). Oosterhof and Todorov (2008) argued that trustworthiness seems to correspond to a judgement of a target's intentions *('do they want to hurt or help you?')* while dominance seemed to index a target's ability to carry out these intentions *('can they hurt or help you?')*. In this model, these two dimensions together represent the evaluation of threat, hypothesised to have an evolutionary background due to the importance of threat perceptions in our ancestral survival (Oosterhof & Todorov, 2008). More recently, Sutherland and colleagues (Sutherland et al., 2013) replicated and extended this model by using highly variable images of faces (so-called 'ambient

images': Burton, Jenkins, & Schweinberger, 2011; Jenkins, White, Van Montfort, & Burton, 2011). This study identified an additional dimension that the researchers called 'youthful-attractiveness', which seemed to link the perception of increasing age with decreasing attractiveness (Sutherland et al., 2013). They suggested that this dimension could represent sexual selection motivations, also with a long evolutionary history.

This evolutionary analysis would suggest that these models represent universal dimensions of impressions made to faces, which would in turn present a powerful pan-cultural aspect of social cognition. Indeed, this is the argument made by Fiske and colleagues for the two key dimensions of warmth and competence that seem to underlie impressions of social groups (Cuddy et al., 2008; Fiske et al., 2007). These warmth and competence dimensions themselves bear a striking resemblance to the trustworthiness (cf. warmth) and dominance (cf. competence) dimensions of facial first impressions described here (e.g. Oosterhof & Todorov, 2008). More generally, Saucier and colleagues (Saucier, Thalmayer, & Bel-Bahar, 2014) examined which concepts used to describe people were universal across languages, and found that morality, competence and attractiveness concepts were common to all twelve languages studied. The semantic differential model of meaning also has three similar dimensions (evaluation, potency and activity) and these conceptual dimensions have also been replicated across cultures (Osgood, 1964; Saucier et al., 2014). Together, these findings indeed suggest that these dimensions may be fundamental aspects of human social cognition that appear across culture.

On the other hand, Oosterhof and Todorov (2008) suggest that facial first impressions are also likely to vary depending on the context and recently, Todorov and colleagues have also explicitly predicted that while some aspects of facial first impressions will be universal, others will be influenced by the culture of the perceiver (Todorov et al., 2015). In order to assess the extent to which the dimensions of facial first impressions are culturally universal, in the current paper, we build models of Chinese perceivers' spontaneous impressions of own- (Asian) and other-culture (Western) faces.

5.2.2. Cross-cultural studies of facial first impressions

A number of previous studies have now examined cross-cultural facial first impressions of traits. In general, this line of research has shown considerable cross-cultural agreement, consistent with models that base facial first impressions in our evolutionary history (Albright et al., 1997; Keating et al., 1981; Rule, Ambady, et al., 2010; Rule, Ishii, & Ambady, 2011; Secord & Bevan, 1956; Walker et al., 2011; Xu et al., 2012; Zebrowitz et al., 2012; Zebrowitz, Montepare, & Lee, 1993; Zebrowitz, Bronstad, et al., 2007). For example, Rule and colleagues (2010) found high agreement between American and Japanese facial trait ratings of politicians; although the traits which accurately predicted votes in real life varied across culture. Rule et al. (2011) also found high cross-cultural agreement for judgements made to the faces of Japanese and American CEOs. Finally, Zebrowitz and colleagues' study (2012) is particularly striking because they showed cross-cultural agreement in trait ratings between American perceivers with perceivers from the Tsimane' people, who live in isolation in the Bolivian rainforest; although the Tsimane' people showed stronger within-culture agreement for own-race compared to American faces. Notable exceptions to this pattern are studies that find that smiling is not always interpreted the same way across cultures (Krys, Hansen, Xing, Szarota, & Yang, 2013; Ozono et al., 2010); or which find evidence for own-race favouritism in the positivity of facial first impressions (Strom, Zebrowitz, Zhang, Bronstad, & Lee, unpublished manuscript; Zebrowitz, Bronstad, et al., 2007). In general, however, studies have largely been supportive of cross-cultural similarity in facial first impressions, if not always in their real-life consequences (Albright et al., 1997; Keating et al., 1981; Secord & Bevan, 1956; Walker et al., 2011; Xu et al., 2012; Zebrowitz et al., 1993), which would seem to suggest that facial impression dimensions may be culturally universal. This also agrees with the findings of a larger literature on the cross-cultural interpretation of emotional facial expressions (e.g. Ekman, 1972, 1994; Elfenbein & Ambady, 2002 although see Jack, Garrod, Yu, Caldara, & Schyns, 2012).

However, these cross-cultural studies were largely designed to target specific hypotheses about particular traits or specified cues to trait judgements, rather than to create culture-specific models or to directly test the universality of the overall structure of facial first impression models. It is clearly valuable to know that perceivers across

cultures can agree on their facial first impressions if directly asked, and establishing the transferability of concepts from one's own culture to another is a useful starting point (J. W. Berry, 1999). This does not however itself substantiate the larger claim that the key dimensions found in Western culture are also the most important dimensions found in other cultures. This is because cross-cultural facial first impression studies have assumed *a priori* which traits are important by following Western models. In order to address this second question, a bottom-up approach is needed to build a model from within a non-Western culture. In the case of Western facial first impressions, Oosterhof and Todorov (2008) took a careful and thorough approach by building an Western model from facial first impressions of Western faces that were spontaneously reported by Western participants. A parallel approach, taken by collecting spontaneous judgements of non-Western faces given by non-Western perceivers, is therefore needed to test whether similar dimensions are found across culture. Without this, research will necessarily constrain the possibility for facial first impression dimensions *other* than Western dimensions to emerge.

5.2.3. East Asian culture

We decided to examine Chinese perceivers' facial first impressions for two reasons. First, examining Chinese perceivers is intrinsically interesting since China is the world's largest country by population, with an estimated 19% of the world's population living there (more than the USA, Oceania and Europe put together: World Population Clock, 2014). Second, examining Chinese perceivers offers a strong test of the universality of facial impression dimensions because there are substantial relevant cultural differences between East Asia and the West (Hofstede, 1980; Kagitcibasi & Berry, 1989; Triandis, 1995). Specifically, East Asian perceivers, especially Chinese perceivers, have been characterised as being more collectivist (having interdependent values and behaviours) than Western participants who have been characterised as being more individualistic (having independent values and behaviours: Hofstede, 1980; Kagitcibasi & Berry, 1989; Nisbett, Peng, Choi, & Norenzayan, 2001; Oyserman, Coon, & Kemmelmeier, 2002; Triandis, 1995). For instance, Chen and colleagues found that (agentic) competence was more important for political voters in the US, whereas social competence was more important to voters in Taiwan (F. F. Chen, Jing, & Lee, 2012; see also Rule, Ambady, et al., 2010). Indeed, competence related-traits may simply have a

stronger social component in East Asian cultures. Yang and Sternberg also report that the concept of intelligence is more linked to social responsibility and social ability in traditional Chinese teaching than in America (S.-Y. Yang & Sternberg, 1997a, 1997b).

Moreover, Wheeler and Kim (1997) found that the content of facial physical attractiveness stereotyping differed for Korean and North American participants, so that physically attractive targets were perceived as higher on traits considered especially important in each culture: that is, collectivist traits for Korean participants, and individualistic traits for North American participants (see also N. Y. Chen, Shaffer, & Wu, 1997; Shaffer, Crepaz, & Sun, 2000; Zebrowitz et al., 1993). This might also lead us to believe that interdependent facial traits are more important to Chinese than Western perceivers than independent facial traits and vice versa.

It might also be possible that judging faces based on individual differences on abstract traits is simply less important to perceivers from a collectivist culture that may instead promote judgements on social or situational grounds. For instance, Morris and Peng (1994) showed that when viewing crime reports, American participants weighted personal factors as more important as explanations for the crime than Chinese participants, and Chinese participants weighted situational factors more highly (Morris & Peng, 1994; see also Lee, Hallahan, & Herzog, 1996; Nisbett et al., 2001). Moreover, multiple studies have found that East Asian participants tended to describe themselves or others in terms of more concrete and less abstract (i.e. trait) descriptions than Western participants, when not given a context (Cousins, 1989; Maass, Karasawa, Politi, & Suga, 2006). This suggests that Chinese perceivers might spontaneously describe fewer individual traits when viewing the faces of strangers than Western perceivers, and instead categorise faces in terms of their social group or mention situational aspects.

5.2.4. Building Chinese facial first impression models

It is therefore currently unknown whether Chinese perceivers spontaneously mention traits as often as Western perceivers, and if so, whether they mention the same types of interpersonal and agentic traits. It is also currently unclear whether the dimensions of Chinese facial first impressions resemble Western ones when judgements are collected

with a bottom-up approach. Answering these questions would add to the cross-cultural psychological literature as well as test which aspects of recent facial first impressions models are culturally universal (e.g. Oosterhof & Todorov, 2008; Sutherland et al., 2013). As others have suggested, there are likely to be both universal and culturally-specific aspects to facial impressions, in agreement with the literature reviewed here (Todorov et al., 2015).

In Study 1, we collected spontaneous judgements made by Chinese perceivers to 60 highly variable images of Asian faces and by British perceivers to 60 highly variable images of Western faces. A free response format allowed us to explore what kinds of facial impressions Chinese perceivers naturally make, as well as minimising the influence of researcher biases or knowledge of previous Western models. It also allowed us to select the most frequently mentioned Chinese facial first impressions, and use these to address our second question in Study 2. In this second study, Chinese perceivers rated a set of 500 highly variable Western and 500 Asian faces for the most frequently mentioned descriptions taken from Study 1. We used factor analysis to build Chinese models of facial first impressions, allowing a strong test of the universality of dimensions of facial first impressions. This allowed us to compare Chinese impressions of Chinese and Western faces in order to examine own-race biases.

5.3. Study 1

Study 1 had two aims. First, we wanted to explore spontaneous impressions made by Chinese perceivers to own-race faces, and to compare these to spontaneous impressions of own-race faces made by Western participants. This allowed us to ascertain whether Chinese and Western perceivers made similar types and frequencies of judgements when seeing an unfamiliar face for the first time. Second, we wanted to use these spontaneous attributions to inform our more ambitious aim of building models of Chinese facial first impressions in Study 2.

In Study 1, Chinese (in China) and British (in the UK) participants freely described a set of 60 own-race faces. We focused on own-race faces for ease of comparison to previous work, which has also used own-race spontaneous descriptions (e.g. Oosterhof &

Todorov, 2008), and in order that the models we built in Study 2 would be built on within-culture impressions. We used 'ambient images' faces taken from the internet; that is, face photographs which are as variable as those encountered in everyday life (Burton et al., 2011; Jenkins et al., 2011). Ambient face images were used so that we could fully sample the range of potential photographic facial cues that Chinese and British perceivers would be reasonably exposed to (for example, while browsing online).

5.3.1. Methods 1

5.3.1.1. Participants 1

Twenty Chinese participants (mean age: 22.0 years, 10 male) were tested in the Guangdong region of Mainland China and 20 British participants (mean age: 23.0 years, 10 male) were tested in York, UK. Participants were all university students; they provided informed consent to procedures that were approved by the ethics committee of the University of York Psychology Department and did not take part in the other currently reported studies. None of the Chinese participants had visited the UK for longer than a month, and likewise none of the British participants had visited China for longer than a month.

5.3.1.2. Facial stimuli

Thirty female and thirty male Western faces were randomly selected from an existing set of 1,000 highly varied 'ambient image' photographs (see Santos & Young, 2005, 2008, 2011). For the current study, we also collected a further 30 female and 30 male adult Asian ambient image faces from the internet. The use of ambient images was pioneered by Burton, Jenkins and Schweinberger (2011; see also Jenkins et al., 2011) in order to stress the potential importance of the variability in everyday images of faces for facial recognition, and ambient images recently been used to examine facial first impressions (Jenkins et al., 2011; Sutherland et al., 2013; Todorov & Porter, 2014; Vernon et al., 2014). Both Asian and Western face sets used in Study 1 therefore were deliberately allowed to vary on many potential facial cues to social impressions, such as pose, expression, lighting and facial paraphernalia such as make-up, hairstyles or glasses (as previous work: see Santos & Young, 2005, 2008, 2011; Sutherland et al., 2013; Vernon et al., 2014 for further details).

The Western set had been previously screened to ensure that the faces represented non-famous Caucasian faces (Santos & Young, 2005, 2008, 2011) and we had seven Asian colleagues look through the Asian image set in order to ensure that all images plausibly represented Chinese faces (since we were testing Chinese perceivers), that they varied as much as the Western faces, and that they only represented non-famous adults. We note that "Asian" is a very broad term encompassing a wide variety of cultures and facial variation, while we had selected the faces to look Chinese in particular. However, for simplicity, we will refer to the face stimuli set as "Asian" rather than Chinese, since we do not know whether the perceivers depicted are indeed Chinese in nationality.

Participants were tested in quiet locations using a laptop computer running a custom Python script (using wxPython toolbox) that could display and collect Unicode characters (i.e. to pick up Simplified Chinese script). Each set of 60 faces was divided into five smaller sets of 12, to ensure that participants were not fatigued when writing down their first impressions and to keep their responses genuine. Face images from one set of these 12 faces were shown to participants one at a time with a blank text box underneath. Participants were told that the study was examining first impressions and were invited to type in anything that came to mind on viewing the face, no matter how silly, judgemental or socially inappropriate. They were encouraged to be honest and reassured that their responses were anonymous and that there was no right or wrong answer. Participants were instructed to write as much as possible until they felt that they were no longer being spontaneous. Instructions, consent forms and demographic questions were translated into Simplified Chinese by a native Chinese speaker and then back-translated (blind) into English by a second native Chinese speaker in order to ascertain that they had the same meaning.

Participants saw 12 experimental faces in a random order, with two additional faces used as a practise (descriptions to these were not analysed). In a second block, participants then saw a different set of 12 faces with a context label. This formed part of a separate study with a different aim and is therefore not analysed or reported here.

5.3.2. Results 1

Our data were the Chinese or British participants' descriptions of the faces (in Simplified Chinese characters or in English), split into units consisting of single words or phrases by one native Chinese and one British researcher (for example, "not friendly"). For ease of comprehension for a Western audience, Chinese descriptions were translated by two native Chinese speakers for presentation here. However, note that the coding and the descriptions chosen for Study 2 were in the original Chinese. That is, participants in Study 2 rated traits that came directly from the participants' descriptions in Study 1.

We found that it was not always possible to separate Chinese descriptions into single concepts since there were several cases where the meaning would change if the concepts were separated. Where one Chinese description meant multiple English words or when different Chinese-English dictionaries disagreed with the exact meaning, we translated the multiple meanings using a forward slash (e.g. "热情" was translated as "passionate/enthusiastic"; note that 热 in isolation could be translated as "warmth" or "heat" while 情 could be translated as "feeling" or "passion"). Where one Chinese description was actually a compound word made up from two original Chinese words we translated it using a tilde (e.g. "和善" was translated as "kind~and~gentle", since this word represents both concepts; note that 和 in isolation means many things, including a union, peace, or "together with"; while 善 means virtuous).

In total, the Chinese perceivers wrote 2,924 characters of Simplified Chinese, which could be divided into 601 units (words and phrases), whereas the British perceivers wrote 3,295 English words, which could be divided into 1,176 units (words and phrases). However, that the Western concepts could be divided into more units does not necessarily mean that the Chinese descriptions were not as rich as the British descriptions, since as discussed above, many Chinese concepts had multiple or compound meanings in English (see also **Discussion 1).**

5.3.2.1. Content of spontaneous descriptions

Five researchers then categorised these descriptions in terms of their content, to fulfil our aim of comparing the type of Chinese and British spontaneous descriptions of faces (see **table 5.1.**). Two native Chinese colleagues worked independently from the original Chinese data and two native English research assistants worked independently from the British data. I also coded both datasets, with the Chinese data translated into English, in order to ensure consistency between the datasets. All coders agreed on their understanding of the categories beforehand. Traits referred to any description of long-lasting character or personality (e.g. "affable", "intelligent", "loves having fun/playful") while emotions referred to feeling states (e.g. "happy", "angry", "quizzically amused"). Appearance words included any description of what the target looked like (e.g. "smile", "haggard", "eyes are glowing", or "cheeky facial expression") but did not include descriptions of the photograph itself (e.g. that it was an old photograph). Sex included both explicit words for gender (e.g. "man", "male", "girl") and also words which implicitly but unambiguously described their gender (e.g. "she", "housewife"). Age likewise included both explicit descriptions of age ("old", "20-30s", "youthful") or implicitly but unambiguously described their age ("lad", "grandmother", "retired"). Explicit and implicit descriptions of age and sex were combined because it proved too difficult to separate these. These five categories were not mutually exclusive: for example, the word 'boy" could be coded as a description involving both age and sex (as in Bond & Cheung, 1983).

Despite potential language differences, there was impressive inter-rater consistency across the three raters: all pairwise kappa values were ≥ .6, well above recommended acceptable thresholds (Bakeman, McArthur, Quera, & Robinson, 1997; Landis & Koch, 1977). Coders then met afterwards and resolved any remaining discrepancies.

Table 5.1. Number of concepts and percentage of coded concepts for Chinese and British perceivers. Coded categories are not mutually exclusive.

	Overall	Trait	Appearance	Sex	Age	Emotion	Total
Chinese	601	30%	20%	14%	10%	3%	66%
British	1178	31%	14%	16%	8%	4%	66%

Table 5.1. presents the types of description given by Chinese and British perceivers. It is clear that the Chinese and British participants produced very similar overall profiles of facial first impressions, with both cultures focusing mostly on targets' traits, then describing their sex and age, and finally describing their appearance (with 49% of the Chinese participants' and 41% of the British participants' appearance descriptions focusing on attractiveness). Interestingly, both cultural groups hardly mentioned the targets' emotions; certainly far less than might be expected, given that this is probably a more objective judgement than some of the trait attributions.

5.3.2.2. Trait descriptions

We then used word clouds (wordle.com) to display the descriptions categorised as involving traits for the Chinese and British participants separately (see **figure 5.1.** see also the **supplementary materials** for the full word clouds for each of the categories). These word clouds present higher frequency descriptions in larger font, and allow exploration of the main concepts emerging from the descriptions (see also Sutherland, Young, Mootz, & Oldmeadow, in press). Following Sutherland and colleagues' (in press) only key words and short phrases (<5 words) were included, to allow examination of the main concepts without common but uninformative words or long phrases hampering this. Three descriptions were removed from the Chinese word cloud, leaving 177 translated trait words and phrases, because the translations could not be fairly summarised (for example; "someone who restrains themselves in appearance but abounds in fervour inside"). For this reason, one description ("critical in a light hearted irrelevant way") was also removed from the British descriptions, leaving 366 trait words and phrases.

It is striking that while the British descriptions cluster around a few main words (mainly "friendly", "kind", "intelligent", "warm" and "nice"), the Chinese descriptions are more variable. However, common themes emerge, with both cultures frequently mentioning traits relating to interpersonal warmth. For instance, the Chinese participants used words translated as "diplomatic", "cheerful/outgoing", "benevolent" and "affable"; while the British participants mentioned "kind", "friendly" and 'warm". Both cultures also mentioned traits perhaps relating to competence, with the Chinese mentioning "capable" and "capable/experienced", and the British mentioning intelligence related traits

("intelligent", "smart" and "wise"). While each culture used culture-specific concepts, this seems to indicate that warmth by competence dimensions might nonetheless underlie trait impressions. Finally, both cultures mentioned seriousness, which might relate to either interpersonal warmth (as standoffish) or competence (as conscientiousness). The Chinese participants also mentioned the word "wretched" which relates to someone being unattractive either in character or appearance (Baidu.com, 2014b).

Figure 5.1. Word clouds depicting first impressions of traits, made to own-culture faces for A) Chinese participants and B) British participants. Larger font size represents more frequent descriptions. (Word clouds from www.wordle.com).

5.3.3. Discussion 1

First, it is striking how similar the types of descriptions made by Chinese and British perceivers were: both cultures mainly mentioned enduring traits, as well as the targets' age, sex and appearance (especially in terms of attractiveness). Crucially, it was clear that Chinese perceivers used trait descriptions as much as the British perceivers did (30% and 31% respectively). This is lower than Oosterhof and Todorov's (2008) report of 68% of free responses to faces by American participants being trait descriptions; perhaps this might be because our study did not code descriptions as mutually exclusive, unlike Oosterhof and Todorov (2008). Interestingly, in the current study, neither cultural group mentioned emotion words very frequently (3% and 4%), which might have been assumed to be a more objective and easily made judgement than the more abstract trait judgements.

It is also worth noting that although the Chinese perceivers wrote less than the British perceivers, it is not necessarily the case that their descriptions were less rich. Often, the Chinese descriptions were more compact but no less informative. In fact, Liao (2014) estimates that Chinese tweets convey around three and a half times as much information as English tweets, although both are limited to 140 characters.

We then focused specifically on participants' descriptions of enduring traits. While most British participants used the same set of trait words ("kind", "friendly" and so forth) there was more variation between Chinese perceivers in trait words they used to describe the faces (**see figure 5.1.**). Nevertheless, strong themes of interpersonal warmth appeared in both the Chinese and British descriptions; which seems to support the idea that warmth traits may be perceivers' primary concern when judging others (Abele & Bruckmüller, 2011; Fiske et al., 2007). However, participants also mentioned competence-based traits, with the British concentrating on intelligence and the Chinese on capability. This seems to indicate that two dimensions of warmth and competence might explain Chinese first impressions of traits, as for Westerners (Oosterhof & Todorov, 2008).

In Study 1 we qualitatively characterised and compared spontaneous descriptions made by Chinese and Western perceivers. Since Study 1 clearly showed that Chinese perceivers do make spontaneous trait inferences, in Study 2 we focused on modelling the key dimensions underlying Chinese facial first impressions. We included the most frequently mentioned trait impressions, along with age, sex and attractiveness since Chinese perceivers also frequently mentioned these attributes (see **table 5.1.**).

5.4. Study 2

Study 2 had three aims. First, we wanted to build models of Chinese facial first impressions. In order to achieve this aim, we chose the 12 most frequently mentioned Chinese first impressions from Study 1 and asked a new set of Chinese participants to rate a highly variable set of 500 Asian and 500 Caucasian faces on these 12 attributes. We then employed factor analysis to reduce the Chinese perceivers' ratings into key dimensions, thereby building separate models for Chinese facial first impressions to Asian faces and to Western faces. By building models from traits that are central to Chinese facial first impressions, the current study therefore offers a strong test of the universality across culture of facial first impression models (Oosterhof & Todorov, 2008; Sutherland et al., 2013).

Second, we wanted to begin a preliminary investigation of how similar Chinese perceptions of traits are to Western perceptions of traits. Since we already have Western ratings on the sample of Caucasian faces (taken from Santos & Young, 2005, 2008, 2011; Sutherland et al., 2013) we could quantify the cross-cultural similarity between Western and Chinese ratings on a variety of traits, by correlating their ratings for Caucasian faces. Since some judgements (e.g. age, sex or attractiveness) are more objectively present in the face than others (e.g. traits: Bruce & Young, 2012), our prediction was that age, sex and attractiveness judgements should have the strongest cross-cultural similarity. Moreover, a wide social psychological literature has repeatedly stressed that the first (trustworthiness or warmth) factor is the most important dimension of social judgement of others (Fiske et al., 2007; Wojciszke, 1994; Wojciszke, Bazinska, & Jaworski, 1998), and that social concepts shows greater cross-cultural stability in meaning than concepts related to skills or accomplishments (Ybarra et al.,

2008). Based on this, we expected that stronger cross-cultural similarity would be found on facial first impressions of warmth compared to facial judgements of competence.

Finally, we wanted to examine the facial judgements made by Chinese perceivers in more detail, including examining the facial cues underlying these judgements. By including both Asian and Caucasian faces, we could examine how these cues differ across facial ethnicity, and any potential own-race facial biases. The own-race bias (or 'cross-race effect') usually refers to the finding that perceivers find it easier to remember individuals from their own racial or cultural group, either due to differential experience with other-race faces, and/or less motivation to individuate other-race faces (e.g. Hugenberg, Young, Bernstein, & Sacco, 2010; Rhodes, Brake, Taylor, & Tan, 1989). Other studies have also found evidence for an own-race bias for emotional expressions, so that emotions are more accurately recognised when perceivers view in-group faces (Elfenbein & Ambady, 2002). Given these previous findings, we might expect that Asian perceivers would find it easier to judge own-race faces, perhaps showing better reliability with their peers for own rather than other race faces. Indeed, Zebrowitz and colleagues found that the Tsimane' people from Bolivia showed higher within-culture agreement when judging Tsimane' faces compared to American faces on their dominance, healthiness and babyfacedness (Zebrowitz et al., 2012; although note that the US raters highly agreed on judgements made to both own- and other-culture faces). Moreover, Zebrowitz and colleagues showed that familiarity mediated trait impressions made by 'White', 'Korean' and 'Black' perceivers to other-race faces, so that in general, more familiar faces were less negatively stereotyped (Zebrowitz et al., 2007; see also Strom et al., unpublished manuscript). We could examine this here by examining the reliability of Chinese ratings for own-culture (Asian) and other-culture (Western) faces.

5.4.1. Methods 2

5.4.1.1. Participants 2

One hundred and twenty-two participants from Mainland China (mean age: 23.58 years, 61 female) volunteered to take part in Study 2 in return for course credit or a small payment. Participants provided informed consent to procedures that were approved by the ethics committee of the University of York Psychology Department. Participants did

not take part in the other currently reported experiments, were recruited and tested at the University of York (UK) and had only been in the UK for an average of 1.47 years (none had been in the UK longer than six years). Before any analyses, two participants were excluded; one reported being of Portuguese nationality and was not raised in Mainland China, and another used her phone while rating and did not pay attention to the experiment.

5.4.1.2. Chinese and Western ambient faces images

For Study 2, we chose 500 Western faces (250 male) at random from an existing set of 1,000 highly varied 'ambient image' photographs (Santos & Young, 2005, 2008, 2011). We also added to the sixty ambient-image Asian faces used in Study 1 by collecting a further 440 Asian faces from the internet, to create a full set of 500 Asian ambient image faces (250 male). Five Asian colleagues (from Mainland China, Hong Kong, Singapore and Korea) examined this larger Asian database, as in Study 1, to make sure that the Asian faces plausibly represented Chinese faces, that they varied as much as the Western faces, and that they only represented non-famous adults. Again, we used ambient image face photographs to ensure that we were sampling a wide range of facial variation and hence potential cues to social judgements; and to allow the most meaningful comparison to the previous finding that three dimensions best explain the variation in Western perceivers' first impressions of naturalistic images (Sutherland et al., 2013; Vernon et al., 2014).

5.4.1.3. Ratings

The 1,000 faces were rated on the most frequently mentioned nine traits from Study 1 (see **table S5.1. in the supplementary material** for the complete list of trait words). Trait frequencies included both positive and negative occurrences of the same root word (e.g. "not intelligent" and "intelligence" were counted as instances of "intelligent") but we did not count multiple occurrences of the same trait word from the same participant in order to avoid basing our choices on idiosyncratic trait use. Ratings of age, sexual dimorphism (femininity-masculinity) and attractiveness were also collected, since age, sex and attractiveness as part of targets' appearance were also frequently mentioned by participants in Study 1 (see **table 5.1.**). Each participant rated all faces on one attribute on a scale of 1 (not very) to 7 (very), for the attributes: 开朗 (cheerful/outgoing), 严肃

（serious), 慈祥 (benevolent), 和蔼 (affable), 和善 (kind~and~gentle), 热情 (passionate/enthusiastic), 干练 (capable/experienced), 圆滑 (diplomatic), 猥琐 (wretched), or 吸引力 (attractive); or from 年轻人 (young) to 老年人(old) or very 女性化的 (feminine) to very 男性化的 (masculine). Note that the traits were rated in Simplified Chinese and that these traits therefore came directly from participants in Study 1, were used in their original form in Study 2, and have only been translated into English here for the convenience of a Western audience.

Participants were tested in a quiet room on a PC computer running PsychoPy software (version 1.76; Peirce, 2007) and were told that they were taking part in a study of first impressions. On each trial, they viewed one face with the rating scale presented underneath. Participants pressed the number key that corresponded with their rating and the next face photograph appeared after a blank interval of approximately 750ms. Participants were given as much time as they wanted but were encouraged to go with their 'gut instinct' (Todorov et al., 2005). As Study 1, instructions, consent forms and demographic questions were translated into Simplified Chinese by a native Chinese speaker and then back-translated (blind) into English by a second native Chinese speaker in order to ascertain that they had the same meaning.

Participants first rated the Asian faces in one block, and then the Western faces in a second block on the same trait. We reasoned that this would offer a better test of genuine Chinese first impressions to Asian faces since, at least until the second block, no mention was made of Western faces or that the study was cross-cultural. We wanted to ensure that participants gave us their genuine first impressions of the faces without explicitly comparing the cultures, at least for the own-race faces (cross-cultural comparison was of course inevitable with the other-race faces). This allows us to more directly compare the own-race face model with previous own-culture models that did not contain any cross-cultural faces (Oosterhof & Todorov, 2008; Sutherland et al., 2013). This also mimics real-life changes in face sampling, where one might move from viewing a predominance of faces from one ethnicity to another as one changes country. The experiment was carried out in Chinese with native Chinese experimenters.

5.4.2. Results 2

5.4.2.1. Reliabilities

We decided *a priori* to collect data from ten participants for each trait, and then allow the reliabilities to vary as a measure of the difference in reliability between own and other culture faces, rather than to keep increasing the sample size until the reliability was acceptable since increasing the number of items (i.e. here, participants) will inevitably increase alpha without implying an increase in quality (Cortina, 1993). Reliabilities were good for the majority of traits for Asian and Western faces (alphas above .7; **see table 5.2.**). However, for judgements of capability/experience, wretchedness, and diplomatic, reliabilities were lower, especially for Western faces (**see table 5.2.**). In general, reliabilities were similar across own- and other-culture faces.

Table 5.2. Reliabilities (Cronbach's alpha) for own and other race facial ratings (n = 10).

	Translation	**Asian faces (own-race)**	**Western faces (other-race)**
热情	Passionate/Enthusiastic	0.93	0.94
开朗	Cheerful/Outgoing	0.89	0.88
严肃	Serious	0.89	0.86
和善	Kind~and~gentle	0.82	0.85
和蔼	Affable	0.81	0.86
慈祥	Benevolent	0.77	0.77
年轻人/老年人	Age	0.96	0.93
猥琐	Wretched	0.63	0.61
女性化的/男性化的	Sexual dimorphism	0.95	0.96
吸引力	Attractive	0.83	0.72
干练	Capable/Experienced	0.66	0.45
圆滑	Diplomatic	0.64	0.41

5.4.2.2. Individual trait ratings

Figure 5.2. displays average face-like images (henceforth referred to as average faces for simplicity) created by averaging together the 20 highest and lowest rated Western and Asian faces for the twelve attributes (using Psychomorph version 6; Tiddeman, Burt, & Perrett, 2001; see Sutherland, 2015 for a practical guide to Psychomorph). Since averaging together faces will remove cues that are inconsistent across faces, these

average faces allow a visualisation of the facial features that consistently cue these attributes.

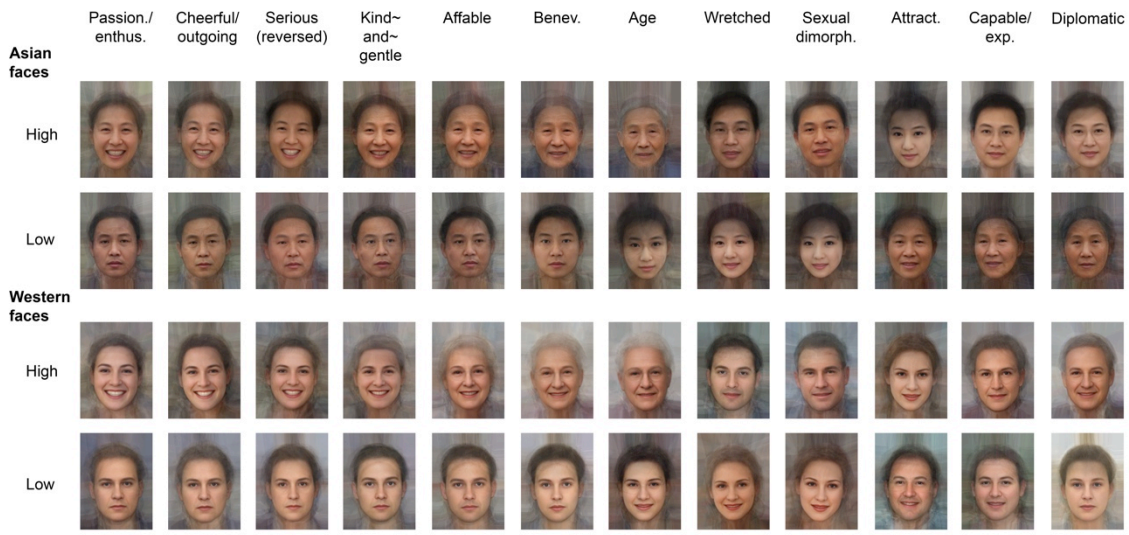

Figure 5.2. Asian and Western face averages made from the 20 faces rated by Chinese perceivers as highest and lowest on passionate/enthusiastic, cheerful/outgoing, seriousness, kind~and~gentleness, affability, benevolence, age, wretchedness, sexual dimorphism, attractiveness, capable/experienced, and diplomatic.

Passionate/enthusiastic, cheerfulness/outgoing, seriousness, and to some extent, kind~and~gentleness, clearly vary on emotional expression, with faces high on these traits displaying a smile (similarly to warmth or trustworthiness for Western perceivers: Oosterhof & Todorov, 2008; Sutherland et al., 2013; Vernon et al., 2014). Affability and benevolence also seem to vary on smiling, along with age. Sexual dimorphism, age and attractiveness seem to rely on cues that have been previously indicated in Western ratings of Western faces, such as skin-tone (darker for males and older adults and less attractive faces), grey hair and wrinkles for older faces, and even skin-tone, femininity and big eyes for attractive faces (Rhodes, 2006; Vernon et al., 2014). Finally, wretchedness appears to vary on the same cues as reversed attractiveness and reversed femininity; and the capable-experienced and diplomatic averages appear to vary on cues which have previously been shown to alter dominance or intelligence (e.g. masculinity and darker skintone: Oosterhof & Todorov, 2008; Sutherland et al., 2013; Vernon et al., 2014; although interestingly, not facial maturity, at least for the Asian faces).

Table 5.3. Correlations between Chinese (vertical) and Western (horizontal) traits for the same 500 Western faces. The Chinese traits were rated by Chinese perceivers and the Western traits by Western perceivers. The largest absolute cross-cultural correlation in each row is in bold.

		Trust.	Warmth	Sexual dimorph.	Age	Attract.	Dom.	Comp.	Intel.
热情	Passionate/ Enthusiastic	.75**	**.91****	-.26**	.02	.21**	-.32**	.08	.20**
开朗	Cheerful/ Outgoing	.73**	**.91****	-.19**	.03	.23**	-.26**	.11*	.21**
严肃	Serious	-.68**	**-.87****	.24**	.02	-.22**	.32**	-.03	-.14**
和善	Kind~and~ gentle	.81**	**.89****	-.26**	.07	.27**	-.32**	.19**	.28**
和蔼	Affable	.83**	**.86****	-.30**	.15**	.24**	-.35**	.21**	.32**
慈祥	Benevolent	.66**	**.69****	-.02	.52**	.01	-.04	.26**	.38**
年轻人/ 老年人	Age	.05	.12**	.34**	**.91****	-.48**	.37**	.13**	.21**
猥琐	Wretched	**-.62****	-.38**	.49**	.05	-.46**	.40**	-.29**	-.36**
女性化的/ 男性化的	Sexual dimorphism	-.42**	-.11*	**.96****	.23**	-.26**	.76**	.15**	-.03
吸引力	Attractive	.26**	.03	-.33**	-.46**	**.66****	-.22**	.31**	.19**
干练	Capable/ Experienced	.18**	.02	.12**	.19**	.28**	.25**	**.51****	.49**
圆滑	Diplomatic	.56**	**.58****	.03	.31**	.21**	.09*	.45**	.45**

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

Table 5.3. depicts the correlations between Chinese and Western perceivers' trait ratings of Western faces. As predicted, across Chinese and Western ratings of Western faces, the more objective judgements of sexual dimorphism ($r = .96$), age ($r = .91$) and attractiveness ($r = .66$) correlated very highly and significantly with Western judgements of the same attributes (see **table 5.3.**). Also as predicted, there was also a cluster of correlations between the warmth traits in both cultures: Chinese ratings of passionate/enthusiastic, cheerfulness/outgoing, kind~and~gentle, seriousness (negatively), affability, and benevolence correlated highly with Western ratings of warmth and also trustworthiness. Chinese ratings of wretchedness correlated (negatively) with Western ratings of both trustworthiness and attractiveness and Chinese ratings of diplomatic correlated highly with both warmth/trustworthiness and competence/intelligence. Finally, Chinese ratings of capable/experienced moderately correlated with Western ratings of competence and intelligence, although it was

interesting that Western ratings of dominance correlated most highly with Chinese ratings of sexual dimorphism rather than capable/experienced (see **table 5.3.**). Overall, these results support our predictions that ratings of sexual dimorphism, age, and attractiveness would be highly correlated across perceiver culture; and that warmth related traits would be more highly correlated than competence related traits across perceiver culture.

5.4.2.3. Chinese model building

Our main aim was to create models of Chinese perceivers' first impressions using factor analysis to reduce the trait impressions into underlying dimensions. We built models for Chinese perceivers' impressions of Asian and Western faces separately. First, Bartlett's test of sphericity indicated that the correlations were large enough that a factor analysis was appropriate for both Asian faces: X^2 (66) = 4847.70, $p < .001$; and Western faces: X^2 (66) = 5429.07, $p < .001$. In order to determine the number of factors to be extracted, an initial principal components analysis was carried out without rotation for the Chinese and Western faces separately, following previous recommendations (Fabrigar et al., 1999). Four criteria were utilised to determine this, in an attempt to be as objective as possible (see Sutherland et al., 2013; Vernon et al., 2014 for further details). For the Asian faces, Kaiser's criterion, a MAP analysis and a parallel analysis all indicated that four factors should be rotated; while the scree test indicated that two, three, four or five factors should be rotated. For the Western faces, Kaiser's criterion, MAP analysis and a parallel analysis all indicated that three factors should be rotated; while the scree test indicated that two, three, four or five factors should be rotated. Since these tests indicated a clear difference in the number of factors across face models, we did not utilise confirmatory factor analysis to test for specific differences across face models (as we have previously carried out for male and female faces in Sutherland, Oldmeadow, & Young, in prep). Most authors recommend that confirmatory factor analyses are only carried out when the number of factors are the same, since otherwise there is no clear starting point (Byrne, 2004).

Having now established the number of factors present for the social attribute ratings of each face set, principal axis factor analyses were run on the Chinese participants' ratings in order to determine the factor structure and loadings (Kline, 1994), with a four

factor solution specified for the Asian faces and a three factor solution specified for the Western faces. A direct oblimin rotation was chosen so that the factors remained oblique; the structure matrix was interpreted, ignoring loadings below .3 (see **table 5.4.**; for further information see Sutherland et al., 2013).

Table 5.4. Chinese principal axis factor analyses for Asian and Western faces: Structure matrices. These can be interpreted as correlations between the factors and variables. Factor loadings ≥ .3 appear in bold.

		Asian face factors				Western face factors		
		1	2	3	4	1	2	3
热情	Passionate/Enthusiastic	**.95**	.12	.29	.10	**.96**	-.15	.07
开朗	Cheerful/Outgoing	**.94**	.06	.22	.11	**.94**	-.10	.08
严肃	Serious	**-.93**	.08	-.21	.03	**-.92**	.12	.03
和善	Kind~and~gentle	**.86**	.20	**.47**	.09	**.93**	-.21	.25
和蔼	Affable	**.75**	**.41**	**.52**	.08	**.90**	-.21	**.37**
慈祥	Benevolent	**.49**	**.78**	**.30**	-.02	**.72**	.17	**.62**
年轻人/老年人	Age	-.13	**.87**	-.20	-.19	.14	**.66**	**.58**
猥琐	Wretched	-.17	-.09	**-.87**	-.21	**-.41**	**.63**	**-.50**
女性化的/男性化的	Sexual dimorphism	**-.31**	.21	**-.53**	.10	-.20	**.51**	-.03
吸引力	Attractive	.13	**-.49**	**.51**	**.44**	.04	**-.68**	.16
干练	Capable/Experienced	-.11	-.13	.05	**.82**	.02	-.13	**.49**
圆滑	Diplomatic	**.32**	-.09	.21	**.74**	**.58**	-.01	**.54**
Alpha (traits ≥.3)		.84	.68	.67	.67	.94	.55	.72

The most striking difference between the Chinese factors for Chinese and Western faces is that the Chinese perceptions of Asian faces needed four factors to best explain the variance, while the Chinese perceptions of the Western faces looked very similar to previous three-dimensional Western models (Sutherland et al., 2013; see **figure 5.3.** for a visualisation of the Chinese factors). A comparable orthogonal PCA (with varimax rotation) found that the four Asian face factors together explained 84% of the variance (factor 1: 37%, factor 2: 17%, factor 3: 15%, factor 4: 15%), while the three Western face factors together explained 77% of the variance (factor 1: 44%, factor 2: 17% and factor 3: 16%).

Nevertheless, it is worth highlighting that while the Chinese model for Asian faces had four dimensions, there were many similarities between this Asian face model and previous Western models (see **figure 5.3.** for a visualisation of the Chinese factors). The first Chinese factor for the Asian faces looked very similar to the approachability or trustworthiness factor described by previous research (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009), with high loadings from passionate/enthusiastic, cheerfulness/outgoing, seriousness (negatively), kind~and~gentleness, affability, and to some extent, benevolence.

The second Chinese factor for the Asian faces looked very similar to the age component of the youthful-attractiveness factor found by our previous study (Sutherland et al., 2013, albeit in reverse) with high loadings from age and benevolence (which in Chinese culture, is strongly age-linked; see the benevolence averages in **figure 5.2.** or Chinese word definitions: Baidu.com, 2015; iciba.com, 2015; MDBG dictionary, 2015). There is a component of decreasing attractiveness to some extent on this age factor too. The third Chinese factor for the Asian faces looked very similar to the attractiveness component of the youthful-attractiveness factor found by our previous study (Sutherland et al., 2013) with high loadings from attractiveness, femininity and (decreasing) wretchedness, an attribute which is strongly linked to interpersonal attractiveness and dating preferences in Chinese culture. “Wretched” describes someone who is creepy, dirty or obscene, unattractive or vulgar in appearance, or insignificant (Baidu.com, 2014b), and the word ‘WSN’ (wěisuǒnán - “wretched man”) is a Pinyin acronym used in popular Chinese culture to describe such a person, usually male (Baidu.com, 2014a; see also the wretchedness averages in **figure 5.2.**). These two culturally-specific perceptions of these attributes may have led to the second and third factors separating, whereas for both the current Chinese ratings of Western faces and for previously collected Western ratings of the same faces, age and attractiveness seemed to be merged (Sutherland et al., 2013).

Finally, the third Chinese factor for the Asian faces looked somewhat similar to previous findings of a dominance factor (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009), with increasing capability/experience, along with diplomacy. Increasing masculinity did not especially contribute to this factor, unlike previous research (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker &

Vetter, 2009) and age contributed negatively to this factor; which is contrary to the current Chinese model of Western faces, and to previous findings regarding the Western dominance factor (Sutherland et al., 2013), at least for male faces (Sutherland et al., in prep).

The first Chinese factor for the Western faces looked very similar to the trustworthiness or approachability factor described by previous research (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009), with high loadings from passionate/enthusiastic, cheerfulness-going, kind~and~gentleness, affability, seriousness (negatively) and benevolence. The second Chinese factor for the Western faces looked very similar to the youthful-attractiveness factor found by our previous study (Sutherland et al., 2013; albeit in reverse) with high loadings from wretchedness, age, masculinity, and (decreasing) attractiveness. Finally, the third Chinese factor for the Western faces looked somewhat similar to previous findings of a dominance or competence factor (Oosterhof & Todorov, 2008; Sutherland et al., 2013, in prep; Walker & Vetter, 2009), with increasing age, diplomacy and capability/experience, although also with high loadings from benevolence, affability and (negatively) wretchedness, which was unexpected. Increasing masculinity did not contribute to this factor, unlike previous research with Western faces and participants (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009).

5.4.2.4. Visualising the Chinese impression models

Figure 5.3. offers a visualisation of the Chinese factors by averaging together the twenty highest and lowest factor scoring faces on each factor using Psychomorph version 6 (Tiddeman et al., 2001; see Sutherland, 2015 for a practical guide to Psychomorph). Factor scores were created using the regression method (see Sutherland et al., 2013 for further details). Briefly, the approachability average face for both Asian and Western face models clearly looks smiling and female, whereas the low average looks male and unsmiling. The high youthful-attractiveness Western average face looks younger, more beautiful and healthier than the low youthful-attractiveness Western average. The high benevolent age Asian face average also looks older and less beautiful than the low benevolent age Asian face average, but is also smiling more; whereas the high attractiveness Asian face looks slightly younger, more attractive and is smiling

more than the low attractiveness Asian face. There is most divergence on cues underlying the last factor for Asian and Western faces, with the high benevolent capability/experience Western face average looking much older, more tanned and perhaps more androgynous than the low Western face counterpart; while the high capability/experience Asian face looks younger, less tanned and more male than the low Asian face counterpart.

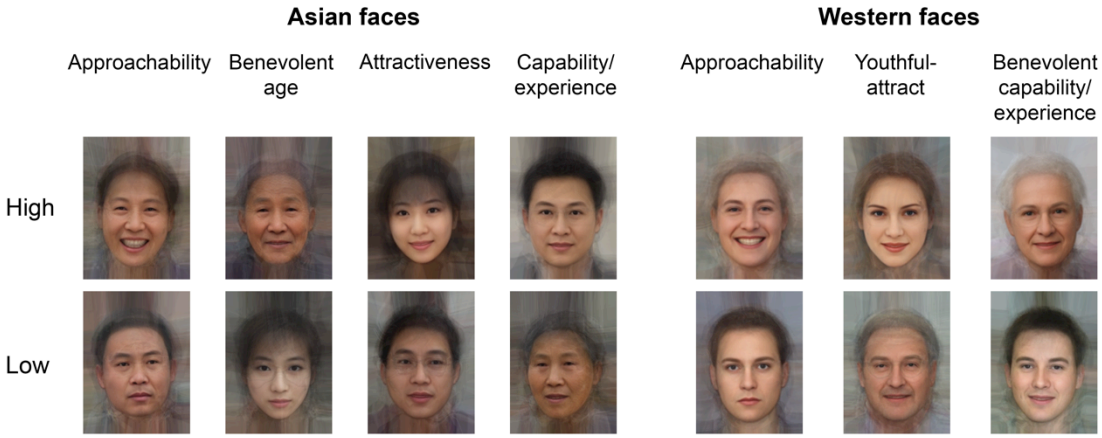

Figure 5.3. High and low face averages for each of the four Chinese and three Western face dimensions, from Chinese perceivers. Each image is an average of the 20 highest or lowest factor-scoring Chinese or Western face photographs for each of the dimensions.

We also ran two-factor solutions for Chinese and Western face ratings, to directly test Oosterhof and Todorov's (2008) two-dimensional model of facial first impressions. For both the Asian and Western faces, the factors seemed to represent approachability and youthful attractiveness (agreeing with the two-dimensional solution for female but not male faces for Western participants; see **Chapter 3**; see **table 5.5.** for the current Chinese factor loadings).

Table 5.5. Asian and Western faces two-dimensional constrained principal axis factor analysis: Structure matrix. These can be interpreted as correlations between the factors and variables. Factor loadings ≥ .3 appear in bold.

		Asian faces		Western faces	
		Factor 1	Factor 2	Factor 1	Factor 2
热情	Passionate/Enthusiastic	**.93**	-.07	**.92**	-.16
开朗	Cheerful/Outgoing	**.88**	-.10	**.90**	-.11
严肃	Serious	**-.83**	.14	**-.86**	.16
和善	Kind~and~gentle	**.91**	-.04	**.93**	-.14
和蔼	Affable	**.83**	.12	**.93**	-.08
慈祥	Benevolent	**.57**	**.49**	**.78**	**.38**
年轻人/老年人	Age	-.07	**.78**	.20	**.85**
猥琐	Wretched	**-.32**	.20	**-.47**	**.33**
女性化的/男性化的	Sexual dimorphism	**-.34**	.25	-.21	**.46**
吸引力	Attractive	.21	**-.74**	.08	**-.50**
干练	Capable/Experienced	-.03	**-.37**	.09	.08
圆滑	Diplomatic	**.36**	**-.40**	**.63**	.18
Alpha (traits ≥.3)		.84	.67	.94	.53

5.4.2.5. Comparing Chinese and Western perceiver models

Finally, we quantified the similarity between the Chinese perceiver model of Western faces depicted here, with a previous Western factor model of Western faces (Sutherland et al., 2013). This was achieved by correlating the factor scores from the current Chinese participant model of Western faces, to factor scores for the same Western faces taken from previous work with Western participants (Sutherland et al., 2013). These correlations are taken at the level of the (500) Western faces. The largest correlation for the first Chinese (approachability) factor was with the first Western (approachability i.e. trustworthiness) factor, $r = .90$, $p < .001$, which is probably close to ceiling given the interrater reliabilities, and since the faces were rated on completely different traits and in different experiments (e.g. in a cross-cultural context for the Chinese participants, and as part of a larger set of 1000 faces for the Western perceivers). The largest correlation for the second (youthful-attractiveness) Chinese factor was with the second (youthful-attractiveness) Western factor ($r = .80$, $p < .001$) and the largest correlation for the third (benevolent capability/experience) Chinese factor was with the third (dominance) Western factor, but while highly significant, was by no means large ($r = .34$, $p < .001$). This gives an indication of the correspondence between the Chinese and

Western factors, and backs up the prediction that the first (trustworthiness or warmth) factor would be the most similar across cultures. However, note that since factor analysis extracts the most robust factors first (i.e. those that explain most of the variance) it is not surprising that later factors correlate less well, since these are based on residual variance after the first factor is extracted (we thank Mike Burton for this point). Indeed, with this in mind, the level of correspondence between the factors across the two cultures is surprisingly high, especially considering the language and trait differences across the perceiver samples.

5.4.3. Discussion 2

In summary, we find that three factors underlie Chinese perceptions of Western faces and that these bear a strong resemblance to Western factor models of Western faces. This resemblance is especially strong for the first and second Chinese factors (approachability, youthful-attractiveness) with the first and second factors in Western models (approachability, youthful-attractiveness: Sutherland et al., 2013). A third Chinese factor (benevolent capability/experience) bore some, but not complete resemblance to a Western dominance factor (Sutherland et al., 2013). However, we also found that Chinese first impressions of own-culture (Asian) faces were more differentiated, with four factors needed to best explain more of the (systematic) variance. These four factors seem to correspond to approachability, benevolent age, attractiveness, and capability/experience. Again, we found the strongest conceptual similarity on the first, approachability factor, as predicted; although the second and third factors also together qualitatively correspond to the youthful-attractiveness factor found in Western research (Sutherland et al., 2013). The separation of these factors may result from culture-specific meanings attached to age (as benevolent) and attractiveness (as not wretched).

5.5. General discussion

It is first worth stressing again that we found evidence for high cross-cultural similarity in both studies. In Study 1, we showed that Chinese perceivers spontaneously make facial first impressions of enduring trait attributes, emotions, age, sex and appearance judgements in very similar proportions to Western perceivers. Moreover, Chinese spontaneous trait attributions, while more variable and of course not using exactly the

same concepts, nevertheless closely echoed Western descriptions, especially regarding the focus in both cultures on interpersonal warmth traits (for example, "cheerfulness/outgoing" or "friendly"). In Study 2, while we found that four factors subserved Chinese perceptions of Asian faces, three factors subserved Chinese perceptions of Western faces and these three factors looked similar to previous research (Sutherland et al., 2013). In general, this shows impressive consistency across cultures (China and the UK) and face groups (Asian and Western) and this provides some support for the suggestion that similar dimensions underpin facial first impressions across cultures.

The strongest cross-cultural similarity (across perceivers and faces) was found on the first, approachability or trustworthiness dimension, which fits with previous findings and theoretical suggestions. For example, Ybarra and colleagues (Ybarra et al., 2008) showed that there is greater cross-cultural stability on the meaning of social traits, relative to agentic, skills based traits. Moreover, Sutherland and colleagues (Sutherland et al., in prep) found that the first, approachability or warmth dimension was highly similar across face gender, but the other two dimensions (dominance or competence and youthful-attractiveness) significantly differed across face gender. A large body of work in social psychology has also suggested and found evidence that the first, warmth or morality dimension is the primary dimension of social cognition (Abele & Bruckmüller, 2011; Cuddy et al., 2008; Wojciszke, 1994; Wojciszke et al., 1998). Interestingly, Rule and collegues (Rule, Ambady, et al., 2010; Rule, Ishii, & Ambady, 2011) found that Japanese and American participants agreed more in terms of their ratings of "power" related facial traits compared to warmth related traits, contrary to the findings in the current study. Possibly this difference can be explained by the fact that Rule and colleagues' (Rule, Ambady, et al., 2010; 2011) studies were carried out in a leadership context and with the faces of real CEOs and politicans, while we did not give participants a context and used a wider range of faces. In future, it would be interesting to see how explicit or implicit contexts may affect cross-cultural similarities and differences.

5.5.1. Cross-cultural differences

While the results were generally fairly supportive of cross-cultural similarity, there were nonetheless some interesting and striking cross-cultural differences. Most importantly, four factors emerged from the Chinese perceivers' impressions of Asian faces. This could reflect the particular emphasis that Chinese perceivers place on old age as having positive social/emotional qualities such as benevolence (Boduroglu, Yoon, Luo, & Park, 2006; Chung & Lin, 2012; Hwang & Han, 2010; Levy & Langer, 1994; Yoon, Hasher, Feinberg, Rahhal, & Winocur, 2000). These perceptions may accurately reflect the fact that older Chinese adults endorse Confucian ethics of benevolence more than younger adults (Lin & Ho, 2009). However, note that other studies find no cross-cultural differences in perceptions of age (Boduroglu et al., 2006; Chan et al., 2012; Cuddy, Norton, & Fiske, 2005; Ryan, Jin, Anas, & Luh, 2004), and one study found that young adults in Hong Kong perceived old age negatively in benevolence (Harwood et al., 1996). In the current study, we deliberately excluded participants from Hong Kong due to the strong British influence on Hong Kong until relatively recently. Many of our participants came from Guangdong (Canton) province in China which remains heavily influenced by traditional Confucian attitudes including respect for one's elders (Chung & Lin, 2012).

A strong relationship was not found for benevolence with the youthful attractiveness factor in the Chinese perceptions of Western faces. Since the Chinese participants in Study 2 had already spent some time in the UK (around one year), perhaps they had learnt that there is less positivity and respect for older age in the UK than in China (Löckenhoff et al., 2009). Our current finding of a difference in the number of factors to own- and other-culture faces may also reflect a form of own-culture bias, with more differentiated first impressions of one's own ethnic group. Alternatively, it is possible that the differences in the number of dimensions between Asian and Western faces simply reflect inherent facial differences. Future work should model British or other Western perceivers' impressions of ambient image Asian faces in order to distinguish between these perceiver and facial explanations.

A second interesting cultural difference was that the Chinese perceivers' capability dimension for both Western and Asian faces seemed rather different to the dominance

dimension proposed by previous models (Oosterhof & Todorov, 2008; Sutherland et al., 2013). There was a strong social component to this dimension; indeed, 'diplomatic' exemplifies social competence and this was a core component to the Chinese capability factor across Asian and Western faces. There were also high loadings from benevolence and affability on this factor for Western faces. This agrees with our previous work, which shows that the dominance or competence dimension seems to vary more across face gender (Sutherland et al., in prep). Moreover, this agrees with studies that have suggested that East Asian (Chinese, Taiwanese and Japanese) conceptions of agentic traits such as intelligence or leadership are more social and moral than Western (in this case, American) conceptions (Hwang, 1999; Jung, Bass, & Sosik, 1995; Rule, Ambady, et al., 2010; S.-Y. Yang & Sternberg, 1997a, 1997b).

5.5.2. The functionality of the dimensions

These findings, along with our previous findings of facial gender differences on the dominance or competence factor (Sutherland et al., in prep) suggest that models of facial first impressions (e.g. Oosterhof & Todorov, 2008; Sutherland et al., 2013) need to take the context into account when examining dimensions of social impression variation. This conclusion agrees with a recent review of facial first impressions (Todorov et al., 2015). One way to resolve an evolutionary account of these dimensions with our current cultural differences may be to focus theoretical discussion and labelling of the dimensions on their underlying functionality, rather than on specific traits.

Specifically, current models of facial and social group perception suggest that the first dimension corresponds to a judgement of a target individual or group's intentions (good or bad), while the other two dimensions correspond to a judgement of the target's capability (ability to carry out their intentions or not) and perhaps of the target's sexual potential (Fiske et al., 2007; Oosterhof & Todorov, 2008; Sutherland et al., 2013). We suggest that these broad intentionality, capability and sexual selection judgements may have evolved either directly or indirectly from mechanisms for judging the threat or opportunities afforded by conspecifics across facial and perceiver culture. However, the relative weightings of cues that are used to judge these broad dimensions are suggested to vary between cultural or social groups, depending on their utility for the context at hand. Social rather than physical dominance cues may be more functional in judging the

capability afforded by Asian faces; for instance, since anger is a cue to dominance (Montepare & Dobish, 2003) and Asian faces are objectively less angry looking (when expressionless) than White faces (Zebrowitz et al., 2010). In addition, because perceivers from East Asian cultures show stronger associations between social and skill based traits such as social skills and intelligence (Li, 2002, 2003; S.-Y. Yang & Sternberg, 1997a, 1997b), or leadership with benevolence (Dickson, Den Hartog, & Mitchelson, 2003; Jung et al., 1995; Rule, Ambady, et al., 2010), these more socially orientated cues may be more functional when Chinese perceivers are making first impressions of capability.

Focusing the theoretical discussion on the underlying functionality of the dimensions may help tie together studies using different trait labels, and draws theoretical attention to the idea that cues or specific traits will change across the context of the judgement made. This allows a more flexible evolutionary analysis of the key dimensions of facial first impressions, which explicitly suggests that ecologically adaptive contextual differences in facial cues will be found. This suggestion is similar to the concept of 'variform universality' in cross-cultural research (Den Hartog, House, Hanges, Ruiz-Quintanilla, & Dorfman, 1999) and with the idea of 'cultural dialects' in facial emotion research (Elfenbein & Ambady, 2002).

5.5.3. Future directions

First, our theoretical analysis of the functional meaning of the dimensions across culture strongly suggests that future studies should explicitly model the effect of the local or global context while measuring perceivers' facial first impressions. For example, given our current findings, we might predict that the dominance dimension is especially variable across different contexts even within a social or cultural group. It would also be interesting to compare perceivers with relatively high or little experience with another cultural context, to ascertain how cultural knowledge impacts facial impressions. For instance, perhaps Western perceivers based in China would perceive age more benevolently.

Second, and perhaps most importantly, here we have only modelled perceptions from one possible cultural group (Chinese perceivers). This was deliberately done in order to

allow a full investigation of first impressions from this group in their own right. We are currently comparing British and other Asian (Hong Kong) perceivers' first impressions of Asian and Western faces in order to deepen comparisons between these two rich cultures. It would also be worth ascertaining whether our current results generalise to groups of perceivers who have never been to the UK or who are isolated from Western culture. While the current study deliberately focused on culture, we have also found evidence for differences across facial gender (Sutherland et al., in prep, in press). In future, it will be interesting to simultaneously examine multiple social groups; for example, to understand whether facial gender differences occur in a cross-cultural context. Finally, within and across countries, human cultural variation is clearly far richer than a simple difference between East and West. An important goal for facial first impressions research in future should be to understand target and perceiver variation in facial first impressions from diverse social groups across the world.

5.6. Conclusions

In the current study we build the first non-Western model of facial first impressions, by first asking Chinese perceivers to provide spontaneous impressions of own-culture faces and then using these to build models of Chinese perceivers' impressions to own-culture (Asian) and other-culture (Western) faces. We found a three-factor structure for Chinese impressions of Caucasian faces, in which the first two factors were very similar to the first two factors of Western perceivers' impressions of Caucasian faces (approachability and youthful-attractiveness: Sutherland et al., 2013). For Asian faces, we found four factors, including an additional attractiveness factor, indicating that there are also culture-specific aspects of facial evaluation. These findings agree with the previous suggestion that both universal and culturally-specific sources of information are used to make first impressions from faces (Todorov et al., 2015). Specifically, here we suggest that while key facial judgements are often based on culturally specific facial cues, these facial judgements similarly function to assess conspecifics' intentions, capability and sexual affordances across culture.

5.7. Supplementary material

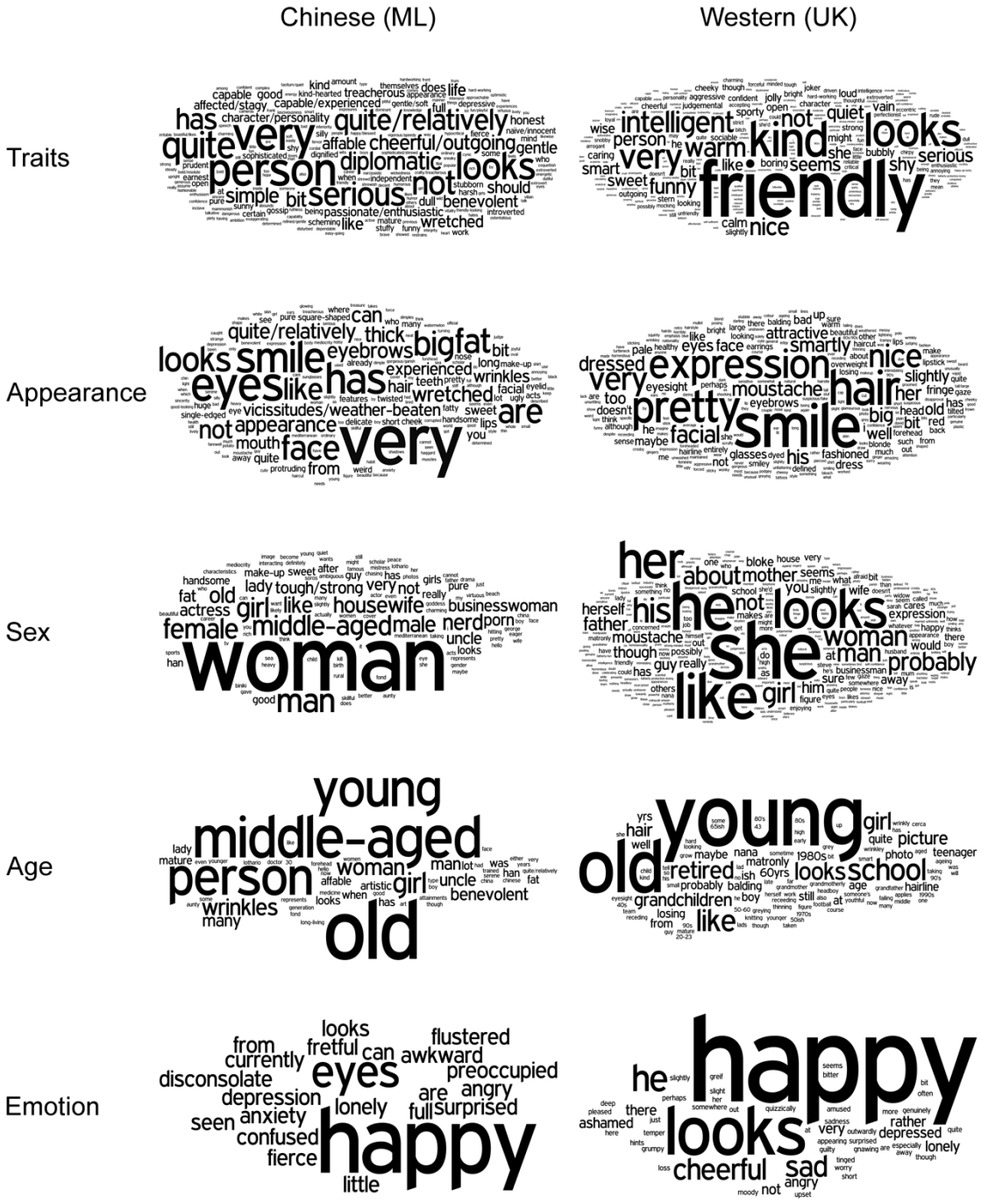

Figure S5.1. Word clouds depicting all descriptions under the five coded categories (age, appearance, emotion, sex and trait categories) for Chinese (translated) and British perceivers in Study 1. Word clouds generated from wordle.com. Note that these word clouds are deliberately allowed to be more inclusive than the word clouds in the main text since they include all words within descriptions that were coded for each category, in order to give the reader an idea of the context around the key words. The only exclusion made was to bar around twenty of the most frequent English words from appearing, such as ‘a’ and ‘the’.

Table S5.1. All trait words mentioned by Chinese participants (left) and British participants (right) from Study 1, sorted by their frequency. Note that here multiple instances of the same words from the same participants are only counted once, so each frequency value represents the number of participants who mentioned each word. Counts include exact antonyms (e.g. "friendly" includes "unfriendly"). The first nine trait labels from Chinese participants were used in Study 2.

Chinese participants (n = 20)			British participants (n = 20)	
Original trait word	Translated trait word	No. of participants	Original trait word	No. of participants
干练	capable/experienced	5	friendly	13
严肃	serious	5	kind	11
慈祥	benevolent	4	intelligent	8
开朗	cheerful/outgoing	4	nice	8
和蔼	affable	3	warm	8
圆滑	diplomatic	3	quiet	7
和善	kind and gentle	3	shy	6
热情	passionate/enthusiastic	3	funny	5
猥琐	wretched	3	sweet	5
做作	affected/stagy	2	bubbly	4
坏坏的	bad	2	calm	4
自信	confident	2	caring	4
抑郁	depressive	2	serious	4
威严	dignified	2	sporty	4
木讷	dull	2	aggressive	3
认真	earnest	2	boring	3
凶	fierce	2	bright	3
滑稽	funny	2	cheeky	3
温柔	gentle/soft	2	confident	3
八卦	gossip	2	extroverted	3
严厉	harsh	2	hard working	3
幽默	humorous	2	joker	3
善良	kind-hearted	2	jolly	3
成熟	mature	2	smart	3
天真	naïve/innocent	2	sociable	3
清纯	pure	2	stern	3
耍心机	scheming	2	vain	3
害羞	shy	2	wise	3
傻	silly	2	annoying	2
固执	stubborn	2	bitch	2
古板	stuffy	2	capable	2
阳光	sunny	2	charming	2
奸诈	treacherous	2	cheerful	2
笑面虎	"friendly-looking villain" [idiom]	1	dull	2
好动	active	1	efficient	2
平易近人	approachable	1	enthusiastic	2
浮夸	boastful/likes	1	fun	2

Chinese participants (n = 20)			British participants (n = 20)	
	exaggerating			
有魄力的	bold/resolute	1	judgemental	2
大胆	brave/bold	1	loyal	2
有一定个性	character/personality [i.e. strong character]	1	mean	2
风情	charming	1	mocking	2
风骚	coquettish	1	naive	2
亲切	cordial	1	open	2
滑头	crafty/treacherous	1	outgoing	2
愤青	cynic	1	rude	2
危险	dangerous	1	snobby	2
正派的	decent	1	strict	2
踏实的	dependable	1	strong	2
坚毅	determined	1	thoughtful	2
强势	dominant	1	tough	2
呆	dull	1	uptight	2
随和	easy-going	1	academic	1
充满活力的	energetic	1	accepting	1
对生活充满积极的	enthusiasm for life	1	affectionate	1
时尚的	fashionable	1	analytical	1
自感良好	feels good about themselves	1	arrogant	1
孝顺	filial piety	1	artistic	1
爽朗	frank and cheerful/outgoing	1	athletic	
友善	friendly	1	authoritative	1
好玩	fun	1	awkward	1
爱玩	fun/playful	1	blunt	1
大方	generous	1	camp	1
幸福	happy/blessed	1	carefree	1
勤勤奋奋的	hardworking/diligent	1	chatty	1
辛苦工作的	hardworking	1	chirpy	1
讨厌外界侵扰的	hates being disturbed	1	cold	1
老实	honest	1	coldly calculating	1
对生活充满希望	hope for life	1	competent	1
虚伪	hypocritical	1	conceited	1
思想犀利	incisive mind	1	condescending	1
独立	independent	1	considerate	1
有自己思想的	independent mind	1	creative	1
自卑	inferiority complex	1	critical	1
内向	introverted	1	crude	1
脾气暴躁	irritable	1	cruel	1
有知识	knowledgeable	1	cutting	1
淫邪	lasciviousness and wickedness	1	daft	1

Chinese participants (n = 20)			British participants (n = 20)	
活泼	lively	1	daydreamer	1
拜金主义	mammonish	1	dedicated	1
自恋	narcissistic	1	dim	1
少讲话	not talkative	1	direct	1
开放	open	1	ditzy	1
乐观	optimistic	1	doer	1
普通的	ordinary	1	down to earth	1
财大气粗	ostentatious	1	dreamy/wistful	1
可怜	pitiful	1	driven	1
人缘不错	popular	1	dry	1
谨慎	prudent	1	earnest	1
斯文	refined/gentle	1	eccentric	1
阅历丰富	rich life experiences	1	egotistical	1
心狠手辣	ruthless	1	excitable	1
严谨	serious and prudent	1	fake	1
精明	shrewd	1	far minded	1
泼辣	shrewish	1	forceful	1
单纯	simple	1	formidable	1
憨厚	simple [or perhaps dull] and honest	1	generous	1
朴实	simple and honest	1	genuine	1
真诚	sincere	1	grave	1
聪明	smart	1	happy-go-lucky	1
鬼鼠	sneaky	1	honest	1
闷骚	someone who restrains themselves in appearance but abounds in fervor inside [hard to translate]	1	imaginative	1
世故	sophisticated	1	inflexible	1
城府颇深	sophisticated/shrewd	1	innocent	1
稳重	steady	1	inquisitive	1
生命力好强	strong vitality	1	insecure	1
沉默	taciturn/quiet	1	lack of common sense	1
正气	upright/a man of integrity	1	lacksidasical	1
雷厉风行的	vigorous/speedy	1	laidback	1
贤惠	virtuous	1	mature	1
有野心	wild ambition	1	meticulous	1
小聪明	wits	1	motherly	1
			narcissism	1
			nasty	1
			nerdy	1
			neutral	1
			noisy	1
			not expressive	1
			not fatherly-protective-loving	1

Chinese participants (n = 20)	British participants (n = 20)	
	not greatly exciting	1
	not shallow	1
	not stupid	1
	not suave	1
	organised	1
	particular	1
	perfectionist	1
	posh	1
	positive	1
	pretentious	1
	prudent	1
	quirky	1
	rebellious	1
	relaxed	1
	reliable	1
	rigid	1
	scatterbrained	1
	scholarly	1
	scrutinising	1
	self-assured	1
	self-centred	1
	self-sufficient	1
	sensible	1
	sharp	1
	showy	1
	sincere	1
	slacker	1
	smarmy	1
	smartypants	1
	snooty	1
	soft	1
	solemn	1
	sour	1
	spiritual	1
	strong-minded	1
	studious	1
	superficial	1
	surly	1
	sympathetic	1
	trustworthy	1
	unassuming	1
	vapid	1
	welcoming	1
	worldly	1
	wry	1

Chapter 6. Discussion

In the last chapter of this thesis, I will first briefly summarise my empirical findings, and then discuss three main themes arising from these. I will then cover potential limitations in the studies carried out here before finishing with a discussion of future directions for the field.

6.1. Summary of findings

In **Chapter 2**, I showed that three dimensions best explained the variance in first impressions of ambient face images (highly varied, naturalistic photographs). While this replicated the trustworthiness and dominance dimensions found by Oosterhof and Todorov (2008), a qualitatively different 'youthful-attractiveness' dimension also emerged, which seemed to combine perceptions of increasing age associated with decreasing health and attractiveness. My explanation for this additional dimension is that this dimension emerged through the inclusion of a larger range of variability in our stimuli images compared to previous research, particularly in terms of age.

In **Chapters 3 and 4**, I examined how categorical face gender fits with these dimensional models of facial first impressions. First, I found that impressions of male and female faces were subserved by slightly different dimensions: male faces showed a clear dominance dimension while female faces instead had a competence dimension (**Chapter 3**). Both genders showed largely similar trustworthiness and youthful-attractiveness dimensions. I also found that male and female faces were spontaneously and explicitly evaluated differently even on the same trait judgements, so that masculinity and dominance was negatively evaluated in female but not male faces (**Chapter 4**).

Finally, in **Chapter 5**, I showed that Chinese participants' spontaneous evaluations of own-race (Asian) faces mirrored spontaneous descriptions of own-race (Western) faces by British participants. Specifically, both cultures frequently mentioned interpersonal traits related to the targets' intentionality (e.g. friendly, kind, benevolent) and competence traits related to the targets' capability (e.g. intelligent, capable/experienced,

wise). Across culture, participants also described concepts related to the traits, age, sex, appearance and emotional expression of the target faces in similar proportions, providing some support for the idea that first impressions of faces may be universal across culture. A second study, which built quantitative models of Chinese ratings of own (Asian) and other (Western) race faces, also found some interesting similarities and differences across face and perceiver culture. For example, Chinese impressions of Western faces showed surprising similarity with previous Western models, especially on the trustworthiness and youthful-attractiveness dimensions. However, Chinese perceptions of Asian faces were best modelled with four dimensions, with approachability, competence and then separate youthful and attractiveness dimensions. This finding perhaps reflects culture-specific attributions of age as 'benevolent' and unattractiveness as 'wretched'.

Together, the findings in **Chapters 2-5** open up a number of themes for discussion, including the value of the ambient face image approach, the need for dimensional accounts of facial first impressions to integrate the influences of social categories, and a discussion of the meaning of the dimensions. These themes will be discussed in detail in the next section.

6.2. Themes

6.2.1. Facial photographic variability: The case for using ambient images

The first theme emerging from this thesis is the suggestion that more face perception studies should be run that include a large, naturalistic range of facial and photographic variability. This builds on seminal recent papers by Jenkins, Burton and colleagues; where they argued that we are ignoring a critical aspect of facial perception by not examining this naturalistic variability between images of the same individuals (Burton et al., 2011; Jenkins et al., 2011). For example, this photographic variability is important in social facial judgements such as trustworthiness or attractiveness, where the variability in social ratings between photographs of the same individual is as least as great as between-individual photographic variability (Jenkins et al., 2011; Todorov & Porter, 2014). In **Chapter 2**, I utilised this concept of ambient images to refer to highly varying images of different individuals (rather than specifically to within-identity

variation) and extended this approach to the investigation of social judgements such as character traits. By modelling judgements made to ambient images for the first time, this revealed that three rather than two dimensions were needed to best explain the variance in social judgements **(see Chapter 2).**

The findings reviewed here suggest that it is important that more face perception studies use ambient images for two reasons. First, as I have shown in this thesis, it is important to ensure that we study a wide and representative range of cues to social impressions in order to fully map the dimensions of social judgement used by perceivers. This is the rationale behind using a large number of variable face photographs, but this research question does not itself require multiple images of a single target individual's face. Second, many of these cues are specific to the photograph used: for instance, differences in facial pose and expression, but also image characteristics such as lighting, background and so forth. To untangle these potential cues to face perception from identity-specific sources of facial variability does require studies with multiple photographs of the same individuals, as in the original ambient image approach (Jenkins et al., 2011). In future studies could further investigate this second point by extending the dimensional approach advanced here to include multiple images of the same individuals, as in the original ambient image study (Jenkins et al., 2011). In particular, it would be interesting to test whether the dimension of approachability/trustworthiness is more variable than the dimension of dominance across different photographs of the same individual. This pattern would be predicted if facial trustworthiness is mainly subserved by emotional expression, which should vary across images of the same person; and if facial dominance is mainly carried by facial structure, which should be relatively stable across images of the same person.

It should be noted that the ambient image theme is not only an appeal for studies to have greater ecological validity, but also a different way of conceptualising face perception in terms of variability rather than as a static judgement or entirely based on face identity. For example, exploring the theme of facial or photographic variability as used in this thesis also opens up the potential for studies to investigate extra-facial cues, such as how the context around the face affects the social perception of traits; from the photographic background to conceptual situations generated by extra-facial labels (see the **Future directions** section). Moreover, the original finding that different face images

of the same people can result in different social judgements has implications for the investigation of the accuracy of social facial judgements. If judgements are accurately perceived from one set of target images, this will not necessarily translate into general accuracy in facial judgements, but may be specific to the photographs used. There is already some evidence for this, since photographs taken during the moment of making a moral decision predict trustworthiness decisions at above chance accuracy, but not photographs of the same individuals taken beforehand or during a practise (Verplaetse et al., 2007). It is also possible that judgements of individuals will be more (or only) accurate when judgements are combined from a range of within-identity photographs of the same individuals.

Finally, to be clear, this theme should not be taken as suggesting that an approach predicated on including high variability in face photographs should replace studies using tightly controlled images. Clearly, there is still value in tightly controlling stimuli face photographs; for example, to increase sensitivity so that even small effects can be found, and to pin-point experimental findings by manipulating cues rather than measuring them. The ambient image approach is thus complementary to the standardised image approach, and potentially, studies can even include both approaches: for instance, **Chapters 2 and 3** also used more tightly controlled average face images. Nevertheless, studies using ambient images may well find different or additional results to studies using tightly controlled images, and a full understanding of face perception will only be achieved by ensuring that the full range of variability in face photographs is used (and indeed, by also using multiple media: see **Limitations** section below).

6.2.2. Dimensions and categories

A second theme explored in this thesis is the issue of understanding how dimensions of facial judgement fit with social categories as perceived from faces. While we might be able to perceive Jane as being more clever but less honest than Beth or John, we can also perceive Jane and Beth as women, and John as a man, and so forth. This theme echoes a recent review of the face perception literature which pointed out that face perception studies mainly examine individual facial recognition, while ignoring insights from social cognition (Quinn & Macrae, 2011). The authors suggest that a full

understanding of face perception will also require an understanding of social categorisation (Quinn & Macrae, 2011).

Across **Chapters 3-5**, I found evidence for individual dimensions of judgement in terms of traits (e.g. trustworthiness, dominance) but that the number **(Chapter 5)**, content **(Chapter 3)** and evaluation of social dimensions **(Chapter 4)** differed across facial social groups. These findings clearly point to the need for models of facial first impressions to integrate social categories with individual dimensions of judgement. Although social groups can be pooled and represented in the exact same space (as was done for male and female faces in **Chapter 2**) by doing this, one ignores important differences in the mechanisms and evaluation of facial impressions. For example, Oosterhof and Todorov (2008) suggest that the second dimension, which they call dominance, is underpinned by cues to masculinity, but this is actually only the case for male faces.

6.2.2.1. Secord's five processes

This theme of integrating categories and dimensions together aligns well with early theoretical descriptions of facial first impressions by Secord in the 1950s (see Secord, 1958, for a review). First, Secord thought that two sets of cues were especially important in making first impressions from faces: 'cultural' cues, which are stereotypes tied to categorical groups such as race, gender and age groups; and 'expressive cues', which are emotional expressions judged from individual faces, especially smiling. This neatly summarises the current facial first impressions literature, including the findings in this thesis. Second, Secord also theorised that traits have to be inferred from facial cues, since facial cues may only ambiguously relate to an enduring trait and since perceivers will search for different traits across situations, depending on their utility, and so a direct one-one mapping between facial cues and every potential trait is unlikely (Secord, 1958). He posited five key inference mechanisms, which are still remarkably useful in summarising the field today.

Briefly, the first process, *'temporal extension'*, refers to the tendency for people to overgeneralise a momentary characteristic, such as a smile, as reflecting an enduring characteristic of the target. This closely relates to the mechanism of emotional

overgeneralisation outlined by later theorists (see the **Introduction** and the next **Theme**). The second process, '*parataxis',* is where a perceiver infers traits from facial cues (or vice versa) based on past experiences with others (Secord, 1958; Zebrowitz & Montepare 2008 also describe a similar process). For example, Lewicki (1985) found that, after initial exposure to a friendly and warm target, participants were more likely to choose a face more closely resembling this target person as being 'kinder and friendlier' than a control face. Evidence for this process was also collected by Secord and Jourard (1956) and more recently, by Verosky and Todorov (2010, 2013). The fourth and fifth processes, *'functional inference'* and *'metaphorical generalization'* refer to the idea that facial cues with functional or metaphorical significance can lead to corresponding trait inferences. For instance, since the lips function for talking, people with large mouths may be seen as talkative, while people with rough or coarse skin may be seen as rough or coarse in personality or behaviour by metaphorical association. Indeed, many traits are themselves metaphors, such as 'warm', 'cold', 'strong-minded', 'bright', 'leonine', or 'foxy'. These two mechanisms have yet to be thoroughly systematically tested, although they were the basis for some of the first empirical studies on facial first impressions (e.g. Thornton, 1944; Zebrowitz, Voinescu, & Collins, 1996), and a recent study has specifically examined metaphorical associations from animal to human faces (Zebrowitz et al., 2011).

Secord's (1958) third process; *'categorization'* contrasts with these previous four processes by representing an inference from a social group rather than individual characteristics of the face. As he outlines: *"In this process, physiognomic cues are utilized to place the person in a category which is associated with certain personality attributes… [For instance] The perceiver estimates the age of the object person from physiognomic cues. Because he believes that older persons are more responsible, more patient, and less energetic, he forms an impression [of this individual] involving these attributes*." (Secord, 1958, p308). This mechanism of (social) categorisation has been relatively overlooked by the field of facial impressions, with important exceptions that will be reviewed in the next section. Of course, the other mechanisms proposed by Secord (1958) and others (Oosterhof & Todorov, 2008) could in some sense also be thought of as categorisation. For example, emotional overgeneralisation could be described as categorisation to the extent that emotions form categories (Young et al., 1997) and these are used to interpret individual signals on the face. Indeed, the dynamic

interactive theory of person construal also describes emotions as categories (Freeman & Ambady, 2011). In the current review, "categorisation" is used to refer only to social group categories, following Secord's (1958) original description.

Importantly, categorisation may help explain the findings in this thesis that the content of key social dimensions differs for faces of different genders and cultures. If perceivers categorise faces in terms of their group membership and then use stereotypes related to these social groups to judge the individual faces, then it follows that key social impressions will differ for faces from different social groups in line with associated stereotypes. My findings in **Chapter 4** extend Secord's (1958) description of the process of categorisation by suggesting that categorisation will also affect how the same facial traits are evaluated, in addition to changing which key traits are inferred.

6.2.2.2. Facial perception evidence for categorisation

Recent face perception studies have also provided evidence for this process of categorisation in facial impressions. For instance, Dotsch and colleagues have found evidence that prejudiced racial associations can predict facial first impressions given to ethnic in- and out-groups (Dotsch, Wigboldus, Langner, & van Knippenberg, 2008; Dotsch, Wigboldus, & van Knippenberg, 2011; for further details see the **Introduction**). The considerable literature by Zebrowitz and others on how perceivers use facial cues which resemble infant characteristics to make inferences about other adults (the babyface stereotype) can also be thought of as an example of categorisation (Zebrowitz, Fellous, Mignault, & Andreoletti, 2003; Zebrowitz & Montepare, 1992; see the **Introduction** for more details).

Secord (1958) also thought that perceivers could also use status, occupation and other social roles to infer traits from faces; and studies have recently shown that perceivers do base some facial trait inferences on stereotypes associated with occupational categories (Imhoff et al., 2013; Oldmeadow et al., 2013). For example, nurses are stereotypically perceived as trustworthy, approachable and healthy, and perceivers' impressions of the extent to which individual faces look like a nurse were associated with others' facial impressions of these same traits (Oldmeadow et al., 2013).

A recent model of person perception, the dynamic interactive theory of person construal (Freeman & Ambady, 2011), also suggests that faces are categorised through facial cues to social groups (such as age or race), and that this leads to associated stereotypical traits being inferred (see the **Introduction** for further details; see also Kunda & Thagard, 1996 for a similar, earlier model). Freeman and Ambady's (2011) model extends Secord's work by suggesting that categorisation is dynamic, with multiple facial cues activating social groups in parallel (e.g. a masculine female face will lead to temporary activations of both 'male' and 'female' neural nodes). They also posit interactive processing from top-down social stereotyping as well as bottom-up visual cues, across all of the model layers. In relation to the current thesis, this suggests that perceivers will activate social categories such as 'male' or 'female' when observing a stranger's face, and this in turn will activate clusters of associated traits while inhibiting unassociated traits. These expectations constrain how the face is perceived. Again, **Chapter 4** extends this model by suggesting that these expectations also constrain how a face is evaluated. Grounding the key dimensions of facial first impressions found in this thesis in the context of Freeman and Ambady's (2011) model helps to provide a theoretical background to the idea that social groups can lead to different patterns of trait activation.

6.2.2.3. Social psychological models of categorisation

Finally, there is also an important body of work in social psychology which has also implicated categorisation as a route to impression formation (Brewer, 1988; Fiske & Neuberg, 1990). However, these models either suggest that categorisation and personalisation (or individualisation) are dual processes (Brewer, 1988) or that they lie on opposite ends of a continuum (Fiske & Neuberg, 1990; see Rapcsak, 2013 for an similar recent model which integrates these ideas with the Bruce and Young 1986 model of face recognition). That is, in these accounts, an impression of a person is either based on a stereotype after they are categorised as a member of a particular social group, or based on a piecemeal analysis of their individual features, which is proposed to be more cognitively demanding (Brewer, 1988; Fiske & Neuberg, 1990; Rapcsak, 2013). The empirical work in this thesis and this review instead suggest that, at least when people are motivated to make an impression of an individual face, people form impressions of individuals based on multiple parallel mechanisms. These mechanisms

include categorisation from social groups, but also mechanisms based on the target's individual facial features such as emotional expression, facial similarity to previous other individuals, or physical maturity (Freeman & Ambady, 2011; Oosterhof & Todorov, 2008; Secord, 1958). Moreover, multiple social categorisations may occur in parallel (Johnson et al., 2012; Johnson & Ghavami, 2011). In other words, I suggest that at least for explicit facial judgements, people form impressions of others as individuals through social categorisation (among other processes), rather than proposing that either individuation or categorisation tends to happen in isolation.

In support of this, when asked to give spontaneous descriptions of individual faces, people frequently mentioned both social roles and individual traits almost interchangeably (see the data in **Chapters 4 and 5;** note that this pattern also appears in Brewer's 1988 own data on descriptions of faces, see p16-17). When mentioning traits, rather than using trait terms categorically, perceivers also frequently used qualifiers to place the individual along a dimension (e.g. "very", "quite", "a bit": see **Chapter 5, supplementary figure S5.1.**). Importantly, these spontaneous impressions of faces were also organised and integrated in terms of the face as an individual person. For example, one Chinese participant described one face as *"Businesswoman, very pretty, very capable. Character/personality is quite gentle/soft"* (translated), while a British participant described another face as *"Boring, depressive but quite intelligent. May [have] done a job such as a solicitor."* Indeed, Brewer's (1988) own data also fit this pattern (p16-17). These observations are best explained by proposing that categorisation into social groups is only one process leading to first impressions of others' faces and that multiple processes integrate to form an impression of an individual (Anderson, 1988; and Rothbart, 1988, provided similar contemporary critiques). However, note that these conclusions refer to person perception as examined in this thesis, where impressions are created sequentially, explicitly and from facial stimuli. The context in which the face is perceived may also change these mechanisms or reduce trait processing of the person (Brewer, 1988; Fiske & Neuberg, 1990; see the **Future directions** section).

As a last point, the dual process and continuum models of person perception (Brewer, 1988; Fiske & Neuberg, 1990) describe the process of categorisation as a way of applying stereotypes to others and this is contrasted to the individualisation process.

However, any impression of a stranger one makes from a facial photograph can be considered to be a stereotype in some sense, even if individualised; since these facial first impressions are based on previous experience with people and the world, rely on limited information, and are not necessarily accurate. This suggestion is compatible with the original description of a stereotype by Lippmann, who included stereotypes of individuals; for example, he described Queen Victoria's image as highly stereotyped in the public mind, along with other celebrities of the day (Lippmann, 1922). Thus, my argument here is that perceivers use social categories as one mechanism to form impressions of others' traits, and to guide evaluation of these impressions. These impressions can be characterised as relating to faces as individuals and in terms of key trait dimensions, yet can also still be thought of as stereotypes.

6.2.3. What do the dimensions themselves mean?

> Sooner or later every psychologist working in the field of personality collides with the problem of trait-names… The nature of his [or her] work forces him [or her] to seek out and to identify dynamic mental structures and sub-structures (habits, needs, sentiments, attitudes, or traits) and to name them. Mathematical symbols cannot be used, for they are utterly foreign to the vital functions with which the psychologist is dealing. Only verbal symbols (ambiguous and troublesome as they are) seem appropriate.
>
> (Allport & Odbert, 1936), *preface page v*

This thesis has dealt throughout with 'trait names' and Allport and Odbert (1936) clearly highlight some of the problems inherent in this. How does one decide the label given to a factor? Throughout this thesis, quantitative evidence was found to justify differences in factor labelling; for example, correlations were carried out between factor scores (**Chapter 5**) and confirmatory factor analysis was used to test factor-loading differences (**Chapter 3**). However, these mathematical processes do not themselves explain the meaning of the dimensions. As Allport and Odbert (1936) point out, this requires a wider theoretical discussion. To address this, I will examine the potential functionality of the dimensions here, and then consider three challenges against this functional account.

6.2.3.1. A functional account of the dimensions

In Oosterhof and Todorov's (2008) original model they suggest that trustworthiness is primarily a judgement of a target's intentions, and dominance is primarily a judgement of a target's capability (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011; Todorov, Said, et al., 2008). Knowing whether a target has good or bad intentions and whether or not they have the capability or not to carry out these intentions, together functions to offer the perceiver an assessment of the target's threat (see the **Introduction**). Oosterhof and Todorov (2008) empirically demonstrated that facial untrustworthiness and dominance together correspond to perceived threat and they argue that these mechanisms have an evolutionary origin in our primate ancestry (Oosterhof & Todorov, 2008). When our evolutionary forebears met conspecifics, if they could rapidly detect the potential threat they posed, they would be more likely to survive or thrive from these interactions. To paraphrase Cheney and Seyfarth, most of the threats encountered by baboons come from other baboons (2007; cited in Todorov & Oosterhof, 2011). Other models of interpersonal and intergroup perception have a similar functional theoretical underpinning to facial first impression models. For instance, the stereotype content model suggests that the two main dimensions of warmth and competence are underpinned by assessments of the intentionality and capability of social groups (Fiske et al., 2007).

6.2.3.2. Challenge one: The finding of three dimensions

The first apparent challenge against Oosterhof and Todorov's (2008) account comes from the finding of an additional youthful-attractiveness dimension in **Chapter 2**. This youthful-attractiveness dimension is not based on a functional analysis of threat and thus appears qualitatively different to the other two dimensions. Nevertheless, since the youthful-attractiveness dimension can be explained by an evolutionary account of sexual selection (i.e. an analysis of the mate potential of conspecifics) it is relatively straightforward to include this extra dimension within the same overarching evolutionary background provided by Oosterhof and Todorov (2008; see **Chapter 2** for further details).

A related criticism is that the functional analysis of the other two dimensions provided by Oosterhof and Todorov (2008) model also focuses too much on descriptions of threat

perception, and therefore relatively neglects the functionality of the positive pole of the trustworthy or dominance/competence dimension. Arguably, in most modern contexts, people do not need to assess the physical threat posed by conspecifics, but instead may rather be judging them on their helpfulness (socially and/or for particular tasks). Of course, an evolutionary mechanism or by-product may not be optimal for modern life. Nevertheless, closer attention to the positive pole of these dimensions in future may help understand how trait judgements are made in everyday contexts such as the workplace or online.

6.2.3.3. Challenge two: Why are social perceptions not always accurate?

The second challenge to a functional model of facial first impressions comes from the finding that trustworthiness and related trait judgements (e.g. honesty) are often inaccurate (see Rule, Krendl, Ivcevic, & Ambady, 2013, and Todorov et al. 2015 for a thorough reviews, and see the **Introduction** for further details). This lack of clear accuracy poses a problem for evolutionary accounts since it is difficult to understand how a trait perception could be selected for if it is not instrumental in promoting survival.

Oosterhof and Todorov (2008) deal with this challenge through the overgeneralisation hypothesis: while character perceptions may not be accurate per se, they are proposed to represent an overgeneralisation of mechanisms that are directly functionally beneficial. For instance, targets with neutral expressions that only slightly resemble emotional expressions may activate genuinely adaptive mechanisms that have developed to judge real emotional expressions (i.e. *emotional overgeneralisation;* see D. S. Berry & McArthur, 1986; Montepare & Dobish, 2003). Thus, genuine (Duchenne) smiling or other emotional cues are selected for, by themselves accurately cuing cooperation or threat in the moment (W. M. Brown, Palameta, & Moore, 2003; Mehu & Dunbar, 2008; Mehu, Grammer, & Dunbar, 2007; Oda, Naganawa, Yamauchi, Yamagata, & Matsumoto-Oda, 2009; Reed, Zeglen, & Schmidt, 2012; Verplaetse et al., 2007). Trustworthiness perceptions then represent, in effect, over-learning by a neural threat system that is ultimately based on emotional expression displays. Similar overgeneralisation mechanisms are proposed for dominance perceptions with other

more veridical facial cues such as facial masculinity or maturity (Oosterhof & Todorov, 2008; Zebrowitz & Montepare, 2008; see **Chapter 3**).

6.2.3.4. Challenge three: Why are the dimensions not fully universal?

A third challenge faced by the functional model of facial first impressions comes from the question of the universality of the dimensions. An evolutionary account of the dimensions would seem to suggest that if the dimensions have functional utility, or at least if they evolved as by-products from functionally useful mechanisms (e.g. judgments of emotional expression) then they should be universal across social and cultural groups. Indeed, Fiske and colleagues explicitly theorise and find evidence for universality for the stereotype content dimensions of group stereotypes (Cuddy et al., 2008; Fiske et al., 2007). Yet, as Todorov and colleagues (2015) have acknowledged, it is likely that facial cues to trait impressions come from both universal and culturally specific sources. In fact, from the evidence in this thesis, while the facial trustworthiness dimension appears (perhaps) surprisingly stable across social groups, there are differences in the other dimensions across gender or cultural groups.

This challenge can be answered by acknowledging that while the facial cues used in overgeneralisation may vary with the population examined, ultimately, the same underlying functional judgements are involved: target intentionality and capability for the Oosterhof and Todorov (2008) and stereotype content models (Fiske et al., 2007). This was tackled in depth in the discussion in **Chapter 3 and Chapter 5** but is summarised briefly again here. To take facial gender differences, human females are naturally less physically strong and less formidable in terms of fighting ability than human males, and structural facial differences between the sexes reflect these differences (Sell et al., 2009). Thus capability is perhaps better judged by competence, and aligned to attractiveness and social status in female faces, rather than dominance, which may be simply a more physical judgement. In addition, while these dimensions might have evolutionary roots that are ultimately derived from selection pressures (e.g. judging a target's capability in order to estimate threat), the mechanisms by which targets are judged as capable are potentially also influenced by cultural norms. For example, East Asian cultural concepts of agentic traits such as intelligence and competence often take on a more social form (Rule, Ambady, et al., 2010; S.-Y. Yang

& Sternberg, 1997a, 1997b). These evolutionary and social accounts are not mutually exclusive, and in fact may be mutually reinforcing. Social contexts that promote intellectual or social skills rather than physical strength may lead to greater visibility of targets with these skills as competent or dominant and thus shift the facial cues used to judge capability overall.

Allowing specific cues to intentionality and capability judgements to be flexible across contexts can also explain why Oosterhof and Todorov's (2008) model of facial perception has a clear dominance dimension, while the stereotype content model of social group perceptions has a dimension of competence (Fiske et al., 2007). It is easier to imagine groups wielding abstract social power (competence or status) than physical power, since with verbal descriptions of groups; one is not immediately faced with physical cues, unlike a face. Certainly, groups and societies have many abstract or social sources of power (for instance, infrastructure, knowledge, ideology, or economic strength: Mann, 1984). Nevertheless, historically and currently, groups or societies which wield physical power, such as military force or manpower, also have more status or capability (Fels, Kremer, & Kronenberg, 2012; Mann, 1984; Marcella, 2004), so it would be interesting to see if photographs rather than verbal descriptions of groups also elicit differences in perceived dominance; and to further investigate the structural arrangement between traits, particularly dominance and competence, for verbally presented social groups (see **Chapter 3**).

While the challenge to universality can be overcome by proposing that different cues lead to the same underlying functionality across contexts, this in turns necessitates that broader labels are used for the dimensions. I propose that the three dimensions found here are called "intentionality", "capability" and "sexual potential" (see **Chapters 3 and 5**). Intentionality and capability reflect the previous discussion on the functional properties of the two main dimensions found in Oosterhof and Todorov's (2008) model. 'Sexual potential' is based on the discussion of the underlying functionality of the third dimension found in **Chapter 2**, on the large literature implicating facial attractiveness for sexual selection (Little et al., 2011; Rhodes, 2006); and on models of partner preferences, which also find a three factor structure underlying person perception, including a dimension composed of sexual attractiveness attributes (Fletcher & Simpson, 2000; Fletcher et al., 1999). This level of description would lift the debate

above the semantics of specific labels while refocusing theoretical attention on the fundamental functionality of these judgements, which can then be further tested. Of course, this is not a new analysis in the sense that the facial evaluation and stereotype content models have already used these functional intentionality/capability explanations (Fiske et al., 2007; Todorov, Said, et al., 2008). Nevertheless, many studies or discussions focus on the trait labels and the dimensions are themselves often labelled by specific traits (Fiske et al., 2007; Oosterhof & Todorov, 2008; Sutherland et al., 2013) and to some extent then the underlying functionality has perhaps been minimised. Moreover, this analysis draws attention to the fact that these underlying relationships have yet to be empirically tested (see **Future directions**).

Refocusing attention on the ultimate functionality of the dimensional labels rather than proximal traits or cues may also help bridge the gap between models of interpersonal and intergroup perception as outlined here, and a more general model of human meaning, the semantic differential (Osgood, 1969; Osgood et al., 1957). This model has three dimensions, evaluation (how good or bad a concept is), potency (how strong or weak a concept is) and activity (how quick or slow a concept is). The semantic differential dimensions are also thought to be the result of evolutionary pressures, with these dimensions having direct functional implications for survival (i.e. one should avoid a tiger, which is bad, strong and quick: Osgood, 1969). Parallels can be drawn between this more general model and the three dimensions found here (see **Chapter 2**) with intentionality reflecting evaluation, capability reflecting potency and sexual potential perhaps replacing activity in the interpersonal domain. An alternative suggestion of course would be to re-label the facial impression dimensions using the semantic differential; however, keeping the focus on intentionality, capability and sexual potential arguably preserves a valuable distinction between the affordances offered by other human agents rather than objects. Moreover, while the first dimension in faces (intentionality) is mostly strongly related to overall positive or negative evaluation (see **Chapter 4**), the other two facial dimensions also contain an evaluative component: competence and physical attractiveness are also positive attributes. This also suggests that ‘intentionality’ rather than ‘evaluation’ better reflects the first dimension as it applies to first impressions of people.

6.2.4. Interim summary

In summary, this section has dealt with the main themes arising from this thesis; the utility of using ambient images of faces to allow for naturalistic variance to be studied and modelled, the need for facial impressions models to integrate categorical social groups and the meaning underlying the dimensions themselves. While understanding or labelling the dimensions of social facial perception is clearly subjective (Allport & Odbert, 1936), I have argued that a functional analysis of the dimensions which grounds them in our evolutionary history may offer considerable theoretical utility. I addressed three challenges to this functional account, and suggested that focusing on the functionality of facial trait perceptions can potentially tie together social and facial models, and that this will help give the facial models enough flexibility to cover judgements made across contexts. In the next section, potential overall limitations of the studies in this thesis will be covered, and then finally the research agenda that follows from this thesis will be set out.

6.3. Potential limitations

In this section, I will address four potential criticisms or limitations of the work presented in this thesis, including the suggestion that these judgements are trivial, the idea that the judgements might reflect the face stimuli samples used rather than a 'ground truth', the criticism that correlations are examined rather than direct experimental manipulations, and finally, that explanations are often based both on stereotypes and evolutionary pathways.

6.3.1. The superficial nature of facial first impressions

Perhaps the most frequent criticism of this field is that often results seem trivial or common sense; for example, it is surely not surprising that people who are smiling look warm and more trustworthy. This is a criticism that has troubled the field of facial first impressions from the start (Secord, 1958). In this important paper, Secord (1958) argued in return that the primary function of social science is to describe and quantify important relationships in the world, not necessarily to discover new relationships. With this in mind, understanding which facial cues give rise to trustworthiness judgements is in itself an important goal, especially given the importance of such judgements in our

social navigation. Such an undertaking becomes even more useful if one can then quantify this relationship and extend predictions to new cases (as we have established in a study using neural network modelling: Vernon, Sutherland, Young, & Hartley, 2014). Secord suggests however that the discovery of new relationships should not be the main goal itself, although if the findings challenge common sense or discover new limits to these relationships, that is, of course, always an interesting and possible consequence (Secord, 1958). Indeed, less obvious results have since been found in the facial first impressions literature; for example, that even the suggestion of a smile on an objectively expressionless face may lead to trustworthiness judgements through emotional overgeneralisation (Oosterhof & Todorov, 2008).

There is an additional important point to make here. To take the example of smiling as a cue to trustworthiness again, it is not necessarily clear that smiling *should* lead to an inference of trustworthiness. For instance, one could imagine a simpler case where smiling leads to an impression of 'happiness' rather than a more enduring trait inference. Yet, when asked to describe first impressions, most descriptions focus on traits (~30%) not emotional feeling (~4%: **see Chapter 5**). It is surely interesting that people use a temporary emotional cue to make an extended prediction about that targets' continuing behaviour at all. As Bruner and colleagues put it: *"In forming impressions of a person…One "knows" more about a person than what seems to be immediately connoted by the acts one has witnessed or the information one has gained about him"* (Bruner, Shapiro, & Tagiuri, 1958, p277). This 'knowledge' of enduring character is suggested to shape future predictions of the target's behaviour, which in turn could shape downstream consequences for how the perceiver behaves towards the target at a subsequent point (Knutson, 1996). Moreover, an impression of an enduring trait such as trustworthiness might have a greater effect on subsequent perceiver behaviour towards the target (for example, if asked to trust them) compared to an impression of a temporary emotional expression such as happiness.

It is also worth highlighting that although I have chosen the salient example of smiling, in actual fact facial cues are highly interconnected and holistically lead to a trait impression (Santos & Young, 2011), so that quite subtle changes can have a large impact on resulting impressions (Vernon et al., 2014). Again, these findings are not obvious or trivial. By changing or controlling cues in isolation, one may affect the

resulting judgement somewhat artificially or even change the nature of the judgement made (see the final **Limitations** section).

Finally, there is sometimes a related misconception that these judgements must be accurate to be interesting to study. At the moment the evidence for the accuracy of facial first impressions is mixed at best (see Rule et al., 2013; Todorov et al., 2015 and the **Introduction** for more details). Importantly, judgements are fairly reliable across raters and moreover, they have potentially profound real-life consequences (see Olivola, Funk, & Todorov, 2014 and the thesis **Introduction**). This makes facial first impressions intrinsically interesting and important to study, regardless of their accuracy.

6.3.2. The use of ambient face photographs

This thesis has limited the examination of facial first impressions to photographs taken from the internet. At best, photographs are an impoverished representation of the cues available during real life encounters; however, they may also be a biased representation. People are likely to deliberately represent themselves in certain ways in photographs compared to what is possible in real-life (e.g. more stereotypically or more attractively). This is all the more important since the ambient image approach samples naturalistic photographs posted online, which may also differ systematically from a wider pool of photographs which are not posted online (for evidence of this selection process, see Hancock & Toma, 2009). In future, it will be interesting to examine what facial characteristics lead people to select photographs across a range of contexts.

Perhaps more importantly, we do not yet know if the dimensions of facial first impressions from static images (as in this thesis) also subserve judgements made from dynamic faces or from 'thin slices' (i.e. 'snapshots' of behaviour: Ambady, Bernieri, & Richeson, 2000; Ambady & Rosenthal, 1992). Arguably, there is a tension between championing an ecological approach to studying face perception with ambient images on the one hand, and limiting the investigation to photographs rather than videos or real life encounters, on the other. This is a reasonable point, and again one that will be interesting to test in the future. If dimensions of social judgements made to static images do correspond to dimensions of social judgements made to videos of faces or thin slices of behaviour, this would emphasise the utility of examining facial

photographs as a proxy for first impressions made in real life. Indeed, a growing number of studies have found a moderate to high correlation between judgements of attractiveness from static photographs and dynamic videos of the same target faces (Kościński, 2013; Lander, 2008; Penton-Voak & Chang, 2008; Rhodes et al., 2011; Roberts et al., 2009; although see also Lander, 2008; Penton-Voak & Chang, 2008; Rubenstein, 2005). However, Gill et al. (2014) showed that judgements of trustworthiness, dominance and attractiveness made to dynamic computer generated faces were reliant on a more complex pattern of multiple emotional expressions than has been described for static faces. For example, trustworthiness judgements were based on cues to surprise as well as happiness (Gill et al., 2014). Moreover, given that judgements of attractiveness and other social attributes are also variable across static photographs of the same people (Jenkins et al., 2011; Todorov & Porter, 2014), there will necessarily be some discrepancy between different sources of social judgements.

Nevertheless, even if judgements made from photographs and videos (or in real life) do not directly correspond, there is still intrinsic theoretical interest in understanding how people make judgements from facial photographs, since people now regularly first meet each other online, where they might be exposed to a single or a few images of each other. Given the increasing influence of the internet in meeting people, whether for business, friendship or dating, understanding how people make first impressions from facial photographs is becoming increasingly important. As already summarised in the **Introduction** of this thesis, social judgements from facial photographs can also predict surprisingly important real-world outcomes, such as the outcome of political elections, online financial lending or court decisions (Olivola et al., 2014).

Finally, one could make this point more specifically about the particular sample of photographs used in this thesis. That is, there may be some artefact about the samples of faces used in the current work. This concern is lessened by the large number of photographs used (500-1,000), but nevertheless, the findings in this thesis would be strengthened by a replication using a different sample of faces. In fact, recently an independent research group investigated the structure of social attribute impressions made to an entirely different sample of controlled faces, and using network analysis, they showed that the traits they examined formed three distinct clusters (Wolffhechel et al., 2014). They named these clusters ‘trustworthiness-friendliness’, ‘attractiveness-

health-extraversion' and 'dominance-masculinity' and drew a parallel between these results and the three dimensions described in this thesis (Wolffhechel et al., 2014). In future, in order to establish the robustness of the dimensional approach, it might be especially useful to examine the generalisation of the dimensions across multiple other face groups (see the **Future directions** section).

6.3.3. Correlation is not causation

In many of the studies in this thesis, correlational relationships are examined, rather than directly manipulating cues and then examining their causal effect on trait judgements. This is both a strength and a weakness of the ambient image approach: on the one hand, this allows multiple interacting cues to be examined together as they naturally co-vary; on the other hand, this correlational approach clearly prohibits causal links to be claimed. Indeed, as I have argued in detail elsewhere (**Chapter 4**), the controlled approach has disadvantages too, since by attempting to isolate facial cues, one arguably changes the nature of the judgement being made. As stated before, for these reasons I believe that the ambient image approach complements rather than replaces a more targeted experimental approach using standardised images. Indeed, I have also employed studies that examine social perceptions from faces after manipulating their facial features with averaging and morphing techniques, which helps substantiate the correlational approach (e.g. Study 1 and 3, **Chapter 2**, Studies 1-2, **Chapter 4**; see also the computer generated faces in Vernon et al., 2014).

Nevertheless, to some extent, issues of causality are unavoidable even with an experimental paradigm, since many of the cues and trait judgements are highly multi-collinear (see **Chapter 2 table S2.1. supplementary material**, see also Vernon et al., 2014). It is therefore hard to isolate one cue or trait judgement without changing others in tandem with the manipulated cue, and this is only made more difficult given the multitude of potential cues or judgements that might be involved. One need only think of the number of words in the English dictionary that refer to traits: from around 4,500 'pure personality traits' to just under 18,000 more general terms used to describe people (Allport & Odbert, 1936). It is therefore almost always possible to argue over which facial cue or judgement was responsible for a particular experimental outcome, even with a carefully controlled experimental study (Todorov et al., 2008 have made a

similar argument). An example of this issue in the literature is the question of whether competence, attractiveness or another facial social judgement is primarily responsible for the finding that facial judgements predict the selection of political candidates (e.g. Olivola & Todorov, 2010; Riggio & Riggio, 2010; Verhulst, Lodge, & Lavine, 2010). Due to this multicollinearity, in some sense, asking which cue is the ultimate cause of a judgement, or which judgement is the ultimate cause of a downstream effect, ends up being somewhat circular. Indeed, the idea of modelling orthogonal dimensions of trait judgements originally arose from a desire to escape an unproductive debate of which specific trait is responsible for an effect, by instead focusing research effort on broad, key dimensions of judgement (Oosterhof & Todorov, 2008; Todorov, Said, et al., 2008). In tandem with this, a more fruitful theoretical approach might be to keep the discussion at the level of broad dimensions, rather than specific traits (see the previous **Themes** section). Similarly, it might be useful to also model key dimensions of physical facial attributes using principal components analysis, and then to relate these facial cue dimensions to the key dimensions of trait judgements.

Finally, a related criticism of studies in this thesis is that where category face or participant groups are manipulated, they are mainly quasi-experiment groups (e.g. gender or cultural groups) rather than having been randomly assigned. To some extent, this was unavoidable since the main focus of this thesis was to examine these social categories, which have often been argued to be primary social categories along with age (e.g. Brewer, 1988; Fiske & Neuberg, 1990; Stangor, Lynch, Duan, & Glas, 1992). However, in future it will be interesting to experimentally manipulate social categories through minimal group paradigms (see the **Future directions** section).

6.3.4. Stereotyping or evolutionary pathway explanations

The theoretical background of this thesis deliberately draws on both evolutionary and stereotyping explanations of findings (e.g. see the **Themes** section) since these explanations are not mutually exclusive, but rather different influences on an outcome or simply different levels of analysis. For instance, social norms representing women as being warm and nice, and men as being competitive and dominant (Prentice & Carranza, 2002) may come from social roles of caregivers and providers (cf. Eagly & Steffen, 1984) but these social roles also reflect the result of evolutionary pressures in our past

(i.e. the biological affordances of males and females, such as physical strength and reproductive abilities: Eagly & Wood, 2013). At a minimum, these traits or roles were not selected against in our evolutionary history. In drawing on both stereotyping and evolutionary mechanisms, I align with previous authors who have attempted to understand evolutionary and social mechanisms of gender and culture simultaneously and in interaction, rather than in isolation (e.g. Eagly & Wood, 2013; Oyserman, Kemmelmeier, & Coon, 2002). Likewise, the use of 'stereotyping' as an explanation of social judgements is used here regardless of the accuracy (and related potential evolutionary function) of these judgements. For example, men are overwhelmingly more likely to be arrested for, convicted of, and admit to committing crimes than women in the UK (Hales, Nevill, Pudney, & Tipping, 2009; Ministry of Justice, 2014), but 'men are more criminal' is nevertheless a stereotype when applied to any individual male facial target.

This thesis also draws on explanations related to both bottom-up visual cues from faces and top-down processing of expectancies attached to social groups. Again, this is deliberate and in line with other authors who have also attempted to build models of social impressions that include both top-down and bottom-up routes and their integration (Anderson, 1988; Freeman & Ambady, 2011; Hassin & Trope, 2000; Quinn & Macrae, 2011). This thesis has established that social groups interact with dimensional cues in faces to create first impressions; in future, more can be done to further investigate the specific routes by which these social group and individual dimensional judgements are affected by stereotyping or other top-down factors (see the **Future directions** section). Understanding these interactions is a priority for face perception research (Quinn & Macrae, 2011). For example, if you make the task competency more stereotypically female, then how does this affect dominance or power evaluations of male and female targets? Similarly, if one spends more time in another culture, do the key dimensions of social face perception shift away from one's own culture towards the surrounding cultural norm?

6.4. Future directions

In this final section, five future directions for the field will be briefly outlined. Specifically, it would be interesting to examine other social groups, to further test the theoretical underpinning of the dimensions, to examine the effect of the context around the face and to establish whether there are individual differences in the perception of facial impressions. Finally, it would be interesting to link behavioural findings with existing neural models.

6.4.1. Other social category groups

First, **Chapter 5** concentrated on Chinese perceiver impressions. However, it would be interesting in future to compare British impressions of Asian and Western faces on own-cultural dimensions with the Chinese impressions. It is possible that four dimensions (including separate age and attractiveness dimensions) would also appear in the British impressions of Asian faces, and that the finding of four dimensions reflects a structural difference between Asian and Western faces rather than a cultural difference between perceivers. Further studies could attempt to cross-validate the cross-cultural similarities and differences found here using a direct rating of similarity (an approach that was used for gender in **Study 4, Chapter 3**). Of course, a full investigation of the universality of facial first impressions also requires many other cultures' spontaneous impressions to be modelled. Setting up a programme of cross-cultural facial impressions research will be an important direction for the field, in order to systematically establish the breadth of the theoretical framework discussed here.

Naturally, this thesis could not examine every social group and there are certainly other social cultural groups that would be interesting to compare. For example, from initial research it seems likely that children and old adults also make similar impressions of faces to young adults (Antonakis & Dalgas, 2009; Boothroyd, Meins, Vukovic, & Burt, 2014; Cogsdill, Todorov, Spelke, & Banaji, 2014; Ewing, Caulfield, Read, & Rhodes, in press; Zebrowitz & Franklin Jr, 2014; Zebrowitz et al., 2013). It would be interesting to know if children or older adults' spontaneous impressions of faces also mirrored young adults' impressions, and whether children or older adults also show the same two- or three-dimensional structure of impressions as young adults (cf. Oosterhof & Todorov, 2008; Sutherland et al., 2013). Another social group to consider in future

research is social class, since British participants spontaneously mentioned social class when viewing faces (e.g. “upper-class”, “posh”; see **Chapter 4, figure 4.3.**) and since social class has a strong influence on real-world outcomes in the UK (Archer, Hutchings, & Ross, 2005; C. White & Edgar, 2010). While research has examined visual cues to social status and power, less is known about visual cues specifically to social class (Fiske & Markus, 2012). In future, studies could explore whether individuals can categorise people as members of different social classes from facial information alone, ascertain whether these facial judgements are accurate, and examine impressions made by perceivers from different social classes.

Finally, it would also be worth experimentally manipulating social categories, since recent work has shown that relatively minimal manipulations can lead to facial processing changing for in-group or out-group members (Cassidy, Quinn, & Humphreys, 2011; Ratner, Dotsch, Wigboldus, van Knippenberg, & Amodio, 2014). Either a minimal group or labelled social group manipulation could be applied to facial first impressions research to further test how associated stereotypes may alter social impressions of photographs of real faces. For example, do verbal in-group labels interact with facial cues to produce an immediate or remembered trait impression?

6.4.2. Further tests of the functionality of the dimensions

Focusing attention on the underlying functionality of the dimensions (see the **Themes** section) highlights further empirical tests that could be carried out to test these ideas. First, facial models have not yet empirically tested the links between trustworthiness with intentionality and dominance with capability. Given the preceding theoretical discussion, one would expect that the judgement of whether a target’s intentions are good or bad should correlate highly with trustworthiness dimension, for instance.

Moreover, behavioural consequences should be linked to intentionality and capability outcomes; that is, people should be more or less willing to actually approach or avoid faces which signal threat (Oosterhof & Todorov, 2008). Indeed, there is now considerable evidence that people avoid economic interactions with faces perceived to be untrustworthy relative to faces perceived to be trustworthy (see Olivola et al., 2014 for a review). Further research could build on this by grounding the social perception of

faces with immediate interpersonal behaviours. For example, virtual reality or masks could be used to manipulate faces on the three social dimensions outlined here and then the physical interpersonal distance between the targets and perceivers could be measured.

6.4.3. Social and contextual factors

Facial first impressions research, along with the majority of face perception research in general, largely examines the cues present in the face without considering the context around the face or the perceiver making the judgement (Quinn & Macrae, 2011; Rule & Tskhay, 2014). Although Oosterhof and Todorov (2008) also state that the context is likely to affect which facial traits are subsequently used in decision making, their model does not explicitly take the context into account by outlining specific contextual predictions. Yet, we never normally perceive people without a context surrounding the face, whether physical (e.g. location, dress and bodily cues) or conceptual (e.g. perceiver goals, stereotyping, social schema). Only by examining faces in context, therefore, will we have a full understanding of face perception more generally (Aviezer et al., 2008; Quinn & Macrae, 2011). The finding that ambient images of the same target across different contexts (photographs) can lead to entirely different facial impressions (Jenkins et al., 2011; Todorov & Porter, 2014) also suggests that contextual differences play a role, and making it timely to investigate how these photographs might differ in appropriateness for different contexts (e.g. at work or for socialising).

It would be interesting to first examine if the presence of a context affects the degree to which people spontaneously make trait inferences at all. One might predict that abstract trait inferences become less influential when a context is included, since perceivers might switch to relatively more concrete attributions, such as social roles or particularly likely behaviour. For example, a couple of studies have found that perceivers mention traits more when describing targets that were verbally presented without a context than within a context (Cousins, 1989; Maass et al., 2006). Second, it would be useful to establish whether the key dimensions of facial first impressions found here and elsewhere (Oosterhof & Todorov, 2008) remain equally important when the context is changed. Oosterhof and Todorov (2008) largely basis their theoretical discussion of the dimensions on an analysis of threat but this does not seem to fully capture what people

usually mean when they talk about first impressions in everyday, usually unthreatening contexts (e.g. the workplace, a restaurant or online). Finally, we might expect the dominance (or competence) dimension to be especially variable across contexts, since one is competent at a task, and thus the judgement of competence surely depends on the task required by the context. This hypothesis could be tested by manipulating the context around the face (e.g. through the photographs chosen or with verbal labels) and examining resulting social perceptions.

6.4.4. Individual differences

In this thesis, I have emphasised the reliability of facial first impressions across raters. Of course, it is clear from the studies shown here and elsewhere that perceivers can agree on impressions (e.g. **Study 1, Chapter 2**) and 'prototypically' trustworthy or dominant faces can even be created which lead to predictable responses from new participants (e.g. **Study 2, Chapter 4**). However, it is also clear that some traits show higher agreement than others (for example, dominance is less reliably judged than trustworthiness: **Chapter 2**). Perceivers also bring their own idiosyncratic judgements to bear; in a broad way, this is demonstrated by the cultural differences found in **Chapter 5**.

In order to address these points, a program of psychometric testing should be established for facial first impressions. This could test how stable individual ratings are over time, how many faces are needed for reliable individual facial and perceiver differences to emerge, and how many individual raters are needed for good predictive validity. This would be especially beneficial in order to establish a standardised clinical test to measure potential disruption in social trait perception of faces, since studies have now started to examine facial first impressions in clinical populations (Blessing, Zöllig, Dammann, & Martin, 2010; Caulfield, Ewing, Burton, Avard, & Rhodes, 2014; Hall et al., 2004, 2012; Haut & MacDonald III, 2010; Mukherjee et al., 2014; Nicol, Pope, Sprengelmeyer, Young, & Hall, 2013; Todorov & Duchaine, 2008; Vakhrusheva et al., 2014; S. White, Hill, Winston, & Frith, 2006).

Individual differences in facial first impressions could also be linked to other stable psychological individual differences. For example, social dominance orientation (SDO)

is the tendency to promote one's group as dominant over other groups (i.e. to prefer stable hierarchies: Pratto, Sidanius, Stallworth, & Malle, 1994; Whitley Jr, 1999). This might affect how one perceives traits from faces belonging to the in-group versus out-groups; for instance, high SDO individuals may especially dislike targets with counterstereotypical faces. It would also be interesting to establish whether facial trait ratings of real images index explicit or implicit prejudice; and if so, whether this could be used as the basis of a test for prejudice. Finally, more general individual differences could be investigated such as Big Five personality dimensions or approach/avoid orientations.

6.4.5. Where and when does social face processing happen in the brain?

Although a full discussion of cognitive neuroscience research on facial first impressions is outside the scope of this thesis, work has started on understanding how these social judgements are processed in the brain (see Bzdok et al., 2011; Mende-Siedlecki, Said, & Todorov, 2013 for recent meta-analyses). For example, neuroimaging (fMRI) research has consistently shown increased activation in the amygdala, and often also in core face processing regions of the brain (i.e. the fusiform gyrus and the superior temporal sulcus) when participants are asked to make facial first impressions or while they view faces chosen to vary on social traits such as trustworthiness (Bzdok et al., 2011; Mende-Siedlecki et al., 2013). In future it will be particularly useful to establish how individual differences or the social or cross-cultural context might alter the neural processing of facial cues in the brain (Beer, Bhanji, Hughes, Freedman, & Fetterolf, 2010; Vuilleumier & Sander, 2008). If differences are found in higher order areas (e.g. prefrontal cortex) this would perhaps suggest that stereotyping is involved, as these regions have been implicated in social cognition studies (e.g. Quadflieg & Macrae, 2011; while differences in visual areas (e.g. the superior temporal sulcus: Mattavelli, Andrews, Asghar, Towler, & Young, 2012) would suggest the involvement of early perceptual differences in facial processing. It will also be interesting to see how fMRI research on facial impressions starts to integrate with classic neural models of face processing in the brain (Haxby et al., 2001; Haxby, Hoffman, & Gobbini, 2000; based on the model of face processing by Bruce & Young, 1986).

In comparison to neuroimaging studies, there has been relatively little research on facial first impressions using electroencephalography (EEG) or event related potentials (ERP) and the research that has been carried out is relatively dissimilar in stimuli, paradigm and in their results (Chiao et al., 2008; Dzhelyova, Perrett, & Jentzsch, 2012; Rudoy & Paller, 2009; D. Yang, Qi, Ding, & Song, 2011). Future studies using EEG could examine the timing of the three broad facial first impressions dimensions rather than specific traits or could directly compare multiple facial first impressions. EEG methods could also be used to test underlying facial impression mechanisms. For example, one elegant ERP method uses the lateralised readiness potential along with a go/no go task as orthogonal tasks onto which perceptual components can be mapped, in order to dissect the order of perceptual processing (for the logic behind this, see Miller & Hackley, 1992; Rahman, van Turennout, & Levelt, 2003). This two-choice method could be used to examine whether judgements of social categories and emotional expression occur before trait impressions such as trustworthiness or dominance, as predicted by facial first impression accounts. Alternatively, this paradigm could be used to test the hypothesis that trustworthiness or warmth judgements are processed before dominance or competence judgements (Abele & Bruckmüller, 2011; Cuddy et al., 2008; Éthier-Majcher, Joubert, & Gosselin, 2013).

6.5. Overall conclusions

In summary, this thesis used ambient image faces to demonstrate that there are three main dimensions of facial first impressions: approachability, youthful-attractiveness and dominance. The first, approachability dimension was especially similar across gender and cultural (British, Chinese) groups; but differences were found between perceiver and facial groups across the other dimensions. Dimensional approaches have considerably advanced the field of face perception by providing an organising framework for research on facial first impressions; however, the findings in this thesis indicate that social category groups need to be integrated into these dimensional models. This can be achieved by returning to an early suggestion that social categorisation is one important mechanism underlying facial first impressions (Secord, 1958), along with other more individualised mechanisms such as emotional and temporal overgeneralisation (Oosterhof & Todorov, 2008).

In future, the priority for facial first impressions research should be to further integrate social psychological insights with face perception to build an overall understanding of how people form impressions of others when they first meet. In particular, it will be interesting to see how the top-down context around the face or individual differences shape facial first impressions, and to systematically map out where and when these facial and social cues are processed in the brain. Finally, the work in this thesis demonstrates the utility of the ambient image approach, which offers more ecological facial stimuli and a wider range of cues for future face perception research.

Chapter 7. References

Abele, A. E., & Bruckmüller, S. (2011). The bigger one of the "Big Two"? Preferential processing of communal information. *Journal of Experimental Social Psychology*, *47*, 935–948.

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.

Adams Jr, R. B., Hess, U., & Kleck, R. E. (2015). The intersection of gender-related facial appearance and facial displays of emotion. *Emotion Review*, *7*(1), 5–13.

Adams Jr, R. B., Nelson, A. J., Soto, J. A., Hess, U., & Kleck, R. E. (2012). Emotion in the neutral face: A mechanism for impression formation? *Cognition & Emotion*, *26*(3), 431–441.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park: Sage.

Albright, L., Malloy, T. E., Dong, Q., Kenny, D. A., Fang, X., Winquist, L., & Yu, D. (1997). Cross-cultural consensus in personality judgments. *Journal of Personality and Social Psychology*, *72*(3), 558–569.

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, Mass.: Addison-Wesley.

Allport, G. W. (1979). *The nature of prejudice* (25th Anniversary Edition). Basic books.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(1), 1–171.

Alrajih, S., & Ward, J. (2014). Increased facial width-to-height ratio and perceived dominance in the faces of the UK's leading business leaders. *British Journal of Psychology*, *105*(2), 153–161.

Amanatullah, E. T., & Tinsley, C. H. (2013). Punishing female negotiators for asserting too much…or not enough: Exploring why advocacy moderates backlash against assertive female negotiators. *Organizational Behavior and Human Decision Processes*, *120*(1), 110–122.

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology*, *32*, 201–271.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–274.

American Society of Plastic Surgeons. (2010). *Report of the 2010 plastic surgery statistics*. Retrieved from http://www.plasticsurgery.org/News-and-Resources/Statistics.html.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology*, *91*(4), 652–661.

Anderson, N. H. (1988). A functional approach to person cognition. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Advances in Social Cognition* (Vol. 1, pp. 37–52). Hillsdale, NJ: Erlbaum.

Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play! *Science*, *323*(5918), 1183.

Archer, L., Hutchings, M., & Ross, A. (2005). *Higher education and social class: issues of exclusion and inclusion*. Routledge.

Arya, A., Jefferies, L. N., Enns, J. T., & DiPaola, S. (2006). Facial actions as visual cues for personality. *Computer Animation and Virtual Worlds*, *17*(3-4), 371–382.

Aviezer, H., Hassin, R., Bentin, S., & Trope, Y. (2008). Putting facial expressions back in context. In *First Impressions* (p. 255). New York: The Guilford Press.

Baidu.com. (2014a). WSN. Retrieved from http://baike.baidu.com/subview/215367/8062859.htm.

Baidu.com. (2014b). 猥琐 (汉语词语) 编辑[wěi suǒ]. Retrieved from http://baike.baidu.com/subview/67302/11149339.htm.

Baidu.com. (2015). 慈祥. Retrieved from http://baike.baidu.com/view/596886.htm.

Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, *2*(4), 357–370.

Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in Cognitive Sciences*, *11*(8), 327–332.

Becker, D. V., Kenrick, D. T., Neuberg, S. L., Blackwell, K. C., & Smith, D. M. (2007). The confounded nature of angry men and happy women. *Journal of Personality and Social Psychology*, *92*(2), 179–190.

Beer, J. S., Bhanji, J. P., Hughes, B. L., Freedman, G., & Fetterolf, J. (2010). The neural architecture of trustworthiness judgments: judging a book by its cover and content. *Frontiers in Neuroscience*, *4*(16), 1–2.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42*(2), 155–162.

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.

Berry, J. W. (1999). Emics and etics: A symbiotic conception. *Culture & Psychology*, *5*(2), 165–171.

Biernat, M. (2012). Stereotypes and shifting standards: Forming, communicating, and translating person impressions. *Advances in Experimental Social Psychology*, *45*, 1–59.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology*, *66*(1), 5–20.

Bijlstra, G., Holland, R. W., & Wigboldus, D. H. J. (2010). The social face of emotion recognition: Evaluations versus stereotypes. *Journal of Experimental Social Psychology*, *46*(4), 657–663.

Blair, I. V., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of Afrocentric features in person perception: judging by features and categories. *Journal of Personality and Social Psychology*, *83*(1), 5–25.

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* (pp. 187–194).

Blessing, A., Zöllig, J., Dammann, G., & Martin, M. (2010). Accurate judgment by dementia patients of neutral faces with respect to trustworthiness and valence. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, *23*(1), 33–38.

Boduroglu, A., Yoon, C., Luo, T., & Park, D. C. (2006). Age-related stereotypes: A comparison of American and Chinese cultures. *Gerontology*, *52*, 324–333.

Bond, C. F., Berry, D. S., & Omar, A. (1994). The kernel of truth in judgments of deceptiveness. *Basic and Applied Social Psychology*, *15*(4), 523–534.

Bond, M. H., & Cheung, T.-S. (1983). College students' spontaneous self-concept: The effect of culture among respondents in Hong Kong, Japan, and the United States. *Journal of Cross-Cultural Psychology*, *14*(2), 153–171.

Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. *Journal of Experimental Psychology: General*, *142*(1), 143–150.

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.

Boothroyd, L. G., Meins, E., Vukovic, J., & Burt, D. M. (2014). Developmental changes in children's facial preferences. *Evolution and Human Behavior*, *35*(5), 376–383.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440.

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, *41*(2), 135–143.

Brewer, M. B. (1988). A dual process model of impression formation. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Advances in Social Cognition* (Vol. 1, pp. 1–36). Hillsdale, NJ: Erlbaum.

Brown, E., & Perrett, D. I. (1993). What gives a face its gender? *Perception*, *22*, 829–840.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. The Guilford Press.

Brown, W. M., Palameta, B., & Moore, C. (2003). Are there nonverbal cues to commitment? An exploratory study using the zero-acquaintance video presentation paradigm. *Evolutionary Psychology: An International Journal of Evolutionary Approaches to Psychology and Behavior*, *1*, 42–69.

Bruce, V., Burton, A. M., Hanna, E., Healey, P., Mason, O., Coombes, A., … Linney, A. (1993). Sex discrimination: how do we tell the difference between male and female face? *Perception*, *22*, 131–152.

Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327.

Bruce, V., & Young, A. W. (2012). *Face perception*. London; New York: Psychology Press.

Bruner, J. S., Shapiro, D., & Tagiuri, R. (1958). The meaning of traits in isolation and combination. In R. Tagiuri & L. Petrullo (Eds.), *Person Perception and Interpersonal Behavior* (pp. 277–288). Stanford, CA: Stanford University Press.

Bull, R., & Hawkes, C. (1982). Judging politicians by their faces. *Political Studies*, *30*(1), 95–101.

Bull, R., Jenkins, M., & Stevens, J. (1983). Evaluations of politicians' faces. *Political Psychology*, 713–716.

Burt, D. M., & Perrett, D. I. (1995). Perception of age in adult Caucasian male faces: computer graphic manipulation of shape and colour information. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *259*(1355), 137–143.

Burton, A. M., Bruce, V., & Dench, N. (1993). What's the difference between men and women? Evidence from facial measurement. *Perception*, *22*, 153–153.

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*, 943–958.

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review*, *100*(2), 204–232.

Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(2), 272–300.

Byrne, B. M. (2009). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Psychology Press.

Bzdok, D., Langner, R., Caspers, S., Kurth, F., Habel, U., Zilles, K., … Eickhoff, S. B. (2011). ALE meta-analysis on facial judgments of trustworthiness and attractiveness. *Brain Structure and Function*, *215*(3-4), 209–223.

Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science*, *20*(10), 1194–1198.

Carré, J. M., Morrissey, M. D., Mondloch, C. J., & McCormick, C. M. (2010). Estimating aggression from emotionally neutral faces: Which facial cues are diagnostic? *Perception*, *39*(3), 356–377.

Cassidy, K. D., Quinn, K. A., & Humphreys, G. W. (2011). The influence of ingroup/outgroup categorization on same-and other-race face processing: The moderating role of inter- versus intra-racial context. *Journal of Experimental Social Psychology*, *47*, 811–817.

Caulfield, F., Ewing, L., Burton, N., Avard, E., & Rhodes, G. (2014). Facial trustworthiness judgments in children with ASD are modulated by happy and angry emotional cues. *PLoS ONE*, *9*(5), e97644.

Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., Löckenhoff, C. E., De Bolle, M., … Terracciano, A. (2012). Stereotypes of age differences in personality traits: Universal and accurate? *Journal of Personality and Social Psychology*, *103*(6), 1050–1066.

Chen, F. F., Jing, Y., & Lee, J. M. (2012). "I" value competence but "we" value social competence: The moderating role of voters' individualistic and collectivistic orientation in political elections. *Journal of Experimental Social Psychology*, *48*(6), 1350–1355.

Chen, N. Y., Shaffer, D. R., & Wu, C. (1997). On physical attractiveness stereotyping in Taiwan: A revised sociocultural perspective. *The Journal of Social Psychology*, *137*(1), 117–124.

Chiao, J. Y., Adams Jr, R. B., Peter, U. T., Lowenthal, W. T., Richeson, J. A., & Ambady, N. (2008). Knowing who's boss: fMRI and ERP investigations of social dominance perception. *Group Processes & Intergroup Relations*, *11*(2), 201–214.

Chiao, J. Y., Bowman, N. E., & Gill, H. (2008). The political gender gap: Gender bias in facial inferences that predict voting behavior. *PLoS ONE*, *3*(10), e3666.

Chung, C., & Lin, Z. (2012). A cross-cultural examination of the positivity effect in memory: United States vs. China. *The International Journal of Aging and Human Development*, *75*(1), 31–44.

Coetzee, V., Re, D., Perrett, D. I., Tiddeman, B. P., & Xiao, D. (2011). Judging the health and attractiveness of female faces: Is the most attractive level of facial adiposity also considered the healthiest? *Body Image*, *8*(2), 190–193.

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

Cohn, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the Facial Action Coding System. *The Handbook of Emotion Elicitation and Assessment*, 203–221.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Cousins, S. D. (1989). Culture and self-perception in Japan and the United States. *Journal of Personality and Social Psychology*, *56*(1), 124–131.

Cracknell, R. (2013). Women in public life, the professions and the boardroom. Commons Library Standard Note. Retrieved from

http://www.parliament.uk/business/publications/research/briefing-papers/SN05170/women-in-public-life-the-professions-and-the-boardroom.

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.

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J. P., … others. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, *48*(1), 1–33.

Cuddy, A. J. C., Norton, M. I., & Fiske, S. T. (2005). This old stereotype: The pervasiveness and persistence of the elderly stereotype. *Journal of Social Issues*, *61*(2), 267–285.

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.

Deaux, K., & Lewis, L. L. (1984). Structure of gender stereotypes: Interrelationships among components and gender label. *Journal of Personality and Social Psychology*, *46*(5), 991–1004.

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

Den Hartog, D. N., House, R. J., Hanges, P. J., Ruiz-Quintanilla, S. A., & Dorfman, P. W. (1999). Culture specific and cross-culturally generalizable implicit leadership theories: Are attributes of charismatic/transformational leadership universally endorsed? *The Leadership Quarterly*, *10*(2), 219–256.

Dickson, M. W., Den Hartog, D. N., & Mitchelson, J. K. (2003). Research on leadership in a cross-cultural context: Making progress, and raising new questions. *The Leadership Quarterly*, *14*(6), 729–768.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*(3), 285–290.

Dotsch, R., & Todorov, A. (2011). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*(5), 562–571.

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.

Dotsch, R., Wigboldus, D. H. J., & van Knippenberg, A. (2011). Biased allocation of faces to social categories. *Journal of Personality and Social Psychology*, *100*(6), 999–1014.

Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *Review of Financial Studies*, *25*(8), 2455–2484.

Dupuis-Roy, N., Fortin, I., Fiset, D., & Gosselin, F. (2009). Uncovering gender discrimination cues in a realistic setting. *Journal of Vision*, *9*(2), 1–8.

Dzhelyova, M. P., Ellison, A., & Atkinson, A. P. (2011). Event-related repetitive TMS reveals distinct, critical roles for right OFA and bilateral posterior STS in judging the sex and trustworthiness of faces. *Journal of Cognitive Neuroscience*, *23*(10), 2782–2796.

Dzhelyova, M. P., Perrett, D. I., & Jentzsch, I. (2012). Temporal dynamics of trustworthiness perception. *Brain Research*, *1435*, 81–90.

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.

Eagly, A. H., & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology*, *5*(1), 1–35.

Eagly, A. H., & Steffen, V. J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. *Journal of Personality and Social Psychology*, *46*(4), 735–754.

Eagly, A. H., & Wood, W. (2013). The nature–nurture debates: 25 years of challenges in understanding the psychology of gender. *Perspectives on Psychological Science*, *8*(3), 340–357.

Ebner, N. C. (2008). Age of face matters: Age-group differences in ratings of young and old faces. *Behavior Research Methods*, *40*(1), 130–136.

Efferson, C., & Vogt, S. (2013). Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports*, *3*(1047), 1–7.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska Symposium on Motivation (1971)* (Vol. 19, pp. 207–282). Lincoln: University of Nebraska Press.

Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychological Bulletin*, *115*(2), 268–287.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203–235.

Etcoff, N. L., Stock, S., Haley, L. E., Vickery, S. A., & House, D. M. (2011). Cosmetics as a feature of the extended human phenotype: Modulation of the perception of biologically important facial signals. *PLoS ONE*, *6*(10), e25656.

Éthier-Majcher, C., Joubert, S., & Gosselin, F. (2013). Reverse correlating trustworthy faces in young and older adults. *Frontiers in Psychology*, *4*(592), 1–7.

Ewing, L., Caulfield, F., Read, A., & Rhodes, G. (in press). Perceived trustworthiness of faces drives trust behaviour in children. *Developmental Science*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/desc.12218/full.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299.

Fels, E., Kremer, J.-F., & Kronenberg, K. (2012). *Power in the 21st Century: International Security and International Political Economy in a Changing World*. Springer Science & Business Media.

Fink, B., Grammer, K., & Thornhill, R. (2001). Human (Homo sapiens) facial attractiveness in relation to skin texture and color. *Journal of Comparative Psychology*, *115*(1), 92–99.

Fink, B., Neave, N., & Seydel, H. (2007). Male facial appearance signals physical strength to women. *American Journal of Human Biology: The Official Journal of the Human Biology Council*, *19*(1), 82–87.

Fiske, S. T. (2012). Managing ambivalent prejudices: Smart-but-cold and warm-but-dumb stereotypes. *The ANNALS of the American Academy of Political and Social Science*, *639*(1), 33–48.

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.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902.

Fiske, S. T., & Markus, H. R. (2012). *Facing social class: How societal rank influences interaction*. Russell Sage Foundation.

Fiske, S. T., & Neuberg, S. L. (1990). A continum of impression formation, from category— based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, *23*, 1–74.

Fiske, S. T., & Taylor, S. E. (1984). *Social Cognition*. Addison-Wesley Pub. Co.

Fiske, S. T., & Taylor, S. E. (1991). *Social Cognition* . Mcgraw-Hill Book Company.

Flannigan, N., Miles, L. K., Quadflieg, S., & Macrae, C. N. (2013). Seeing the unexpected: Counterstereotypes are implicitly bad. *Social Cognition*, *31*(6), 712–720.

Fletcher, G. J. O., & Simpson, J. A. (2000). Ideal standards in close relationships: their structure and functions. *Current Directions in Psychological Science*, *9*(3), 102–105.

Fletcher, G. J. O., Simpson, J. A., & Thomas, G. (2000). Ideals, perceptions, and evaluations in early relationship development. *Journal of Personality and Social Psychology*, *79*(6), 933–940.

Fletcher, G. J. O., Simpson, J. A., Thomas, G., & Giles, L. (1999). Ideals in intimate relationships. *Journal of Personality and Social Psychology*, *76*(1), 72–89.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–279.

Freeman, J. B., Johnson, K. L., Ambady, N., & Rule, N. O. (2010). Sexual orientation perception involves gendered facial cues. *Personality and Social Psychology Bulletin*, *36*, 1318–1331.

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). Looking the part: Social status cues shape race perception. *PLoS ONE*, *6*(9), e25107.

Friedman, H., & Zebrowitz, L. A. (1992). The contribution of typical sex differences in facial maturity to sex role stereotypes. *Personality and Social Psychology Bulletin*, *18*, 430–438.

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.

Gart, J. J., & Zweifel, J. R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika*, *54*(1/2), 181–187.

Geniole, S. N., Molnar, D. S., Carré, J. M., & McCormick, C. M. (2014). The facial width-to-height ratio shares stronger links with judgments of aggression than with judgments of trustworthiness. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1526–1541.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, *60*(4), 509–517.

Gill, D., Garrod, O. G., Jack, R. E., & Schyns, P. G. (2014). Facial movements strategically camouflage involuntary social signals of face morphology. *Psychological Science*, *25*(5), 1079–1086.

Gilmore, D. C., Beehr, T. A., & Love, K. G. (1986). Effects of applicant sex, applicant physical attractiveness, type of rater and type of job on interview decisions. *Journal of Occupational Psychology*, *59*(2), 103–109.

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*(3), 491–512.

Glick, P., & Fiske, S. T. (1997). Hostile and benevolent sexism measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly*, *21*(1), 119–135.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168.

Gunn, D. A., Rexbye, H., Griffiths, C. E. M., Murray, P. G., Fereday, A., Catt, S. D., … et al. (2009). Why some women look young for their age. *PLoS ONE*, *4*(12), e8021.

Hales, J., Nevill, C., Pudney, S., & Tipping, S. (2009). *Longitudinal Analysis of the Offending, Crime and Justice Survey 2003-06*. Home Office. Retrieved from https://www.gov.uk/government/publications/the-offending-crime-and-justice-survey-longitudinal-analysis-2003-to-06.

Hall, J., Harris, J. M., Sprengelmeyer, R., Sprengelmeyer, A., Young, A. W., Santos, I. M., … Lawrie, S. M. (2004). Social cognition and face processing in schizophrenia. *The British Journal of Psychiatry*, *185*(2), 169–170.

Hall, J., Philip, R. C. M., Marwick, K., Whalley, H. C., Romaniuk, L., McIntosh, A. M., … Stanfield, A. C. (2012). Social cognition, the male brain and the autism spectrum. *Plos ONE*, *7*(12), e49033.

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.

Harrington, D. (2008). *Confirmatory factor analysis*. Oxford University Press, USA.

Harwood, J., Giles, H., Ota, H., Pierson, H. D., Gallois, C., Ng, S. H., … Somera, L. (1996). College students' trait ratings of three age groups around the Pacific Rim. *Journal of Cross-Cultural Gerontology*, *11*(4), 307–317.

Hassin, R., & Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology*, *78*(5), 837–852.

Haut, K. M., & MacDonald III, A. W. (2010). Persecutory delusions and the perception of trustworthiness in unfamiliar faces in schizophrenia. *Psychiatry Research*, *178*(3), 456–460.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

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.

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.

Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks?: The implied communality deficit. *Journal of Applied Psychology*, *92*(1), 81–92.

Hess, U., Adams Jr, R. B., & Kleck, R. E. (2004). Facial appearance, gender, and emotion expression. *Emotion*, *4*(4), 378–388.

Hess, U., Adams Jr, R. B., & Kleck, R. E. (2005). Who may frown and who should smile? Dominance, affiliation, and the display of happiness and anger. *Cognition & Emotion*, *19*(4), 515–536.

Hill, H., Bruce, V., & Akamatsu, S. (1995). Perceiving the sex and race of faces: The role of shape and colour. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *261*(1362), 367–373.

Hilton, J. L., & Von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*(1), 237–271.

Hochschild, T. R., & Borch, C. (2011). About face: The association between facial appearance and status attainment among military personnel. *Sociological Spectrum*, *31*(3), 369–395.

Hofstede, G. (1980). Culture and organizations. *International Studies of Management & Organization*, 15–41.

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(2), 199–209.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Hosoda, M., Stone-Romero, E. F., & Coats, G. (2003). The effects of physical attractiveness on job-related outcomes: A meta-analysis of experimental studies. *Personnel Psychology*, *56*(2), 431–462.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice. *Psychological Science*, *14*(6), 640–643.

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science*, *15*(5), 342–345.

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: An integrative account of the other-race recognition deficit. *Psychological Review*, *117*(4), 1168–1187.

Hugenberg, K., Young, S. G., Sacco, D. F., & Bernstein, M. J. (2011). Social categorization influences face perception and face memory. In A. J. Calder, G. Rhodes, & M. Johnson (Eds.), *Oxford Handbook of Face Perception* (pp. 245–262). Oxford: Oxford University Press.

Hwang, K. (1999). Filial piety and loyalty: Two types of social identification in Confucianism. *Asian Journal of Social Psychology*, *2*(1), 163–183.

Hwang, K., & Han, K. (2010). Face and morality in Confucian society. In M. H. Bond (Ed.), *The Oxford Handbook of Chinese Psychology* (pp. 479–498). Oxford University Press.

Iciba.com. (2015). 慈祥. Retrieved from http://www.iciba.com/慈祥.

Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R. (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology*, *4*(386), 1–8.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241–7244.

Jaquet, E., & Rhodes, G. (2008). Face aftereffects indicate dissociable, but not distinct, coding of male and female faces. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 101–112.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*, 313–323.

Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology*, *102*(1), 116–131.

Johnson, K. L., & Ghavami, N. (2011). At the crossroads of conspicuous and concealable: What race categories communicate about sexual orientation. *PLoS ONE*, *6*(3), e18025.

Jones, A. L., Kramer, R. S. S., & Ward, R. (2012). Signals of personality and health: The contributions of facial shape, skin texture, and viewing angle. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1353–1361.

Jung, D. I., Bass, B. M., & Sosik, J. J. (1995). Bridging leadership and culture: A theoretical consideration of transformational leadership and collectivistic cultures. *Journal of Leadership & Organizational Studies*, *2*(4), 3–18.

Kagitcibasi, C., & Berry, J. W. (1989). Cross-cultural psychology: Current research and trends. *Annual Review of Psychology*, *40*(1), 493–531.

Keating, C. F., Mazur, A., & Segall, M. H. (1981). A cross-cultural exploration of physiognomic traits of dominance and happiness. *Ethology and Sociobiology*, *2*(1), 41–48.

Keefe, B. D., Dzhelyova, M., Perrett, D. I., & Barraclough, N. E. (2013). Adaptation improves face trustworthiness discrimination. *Frontiers in Psychology*, *4*(358), 1–7.

Kleisner, K., Chvátalová, V., & Flegr, J. (2014). Perceived intelligence is associated with measured intelligence in men but not women. *PLoS ONE*, *9*(3), e81237.

Kline, P. (1994). *An easy guide to factor analysis*. Psychology Press.

Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*, *20*(3), 165–182.

Kościński, K. (2013). Perception of facial attractiveness from static and dynamic stimuli. *Perception*, *42*(2), 163–175.

Kramer, R. S. S., Jones, A. L., & Sharma, D. (2013). Sequential effects in judgements of attractiveness: The influences of face race and sex. *PloS One*, *8*(12), e82226.

Kramer, R. S. S., & Ward, R. (2010). Internal facial features are signals of personality and health. *The Quarterly Journal of Experimental Psychology*, *63*(11), 2273–2287.

Kramer, R. S. S., & Ward, R. (2011). Different signals of personality and health from the two sides of the face. *Perception*, *40*(5), 549–562.

Krings, F., Sczesny, S., & Kluge, A. (2011). Stereotypical inferences as mediators of age discrimination: The role of competence and warmth. *British Journal of Management*, *22*(2), 187–201.

Krys, K., Hansen, K., Xing, C., Szarota, P., & Yang, M. (2013). Do only fools smile at strangers? Cultural differences in social perception of intelligence of smiling individuals. *Journal of Cross-Cultural Psychology*, *45*(2), 314–321.

Kunda, Z., Sinclair, L., & Griffin, D. (1997). Equal ratings but separate meanings: Stereotypes and the construal of traits. *Journal of Personality and Social Psychology*, *72*(4), 720–734.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*(2), 284–308.

Lammers, J., Gordijn, E. H., & Otten, S. (2009). Iron ladies, men of steel: The effects of gender stereotyping on the perception of male and female candidates are moderated by prototypicality. *European Journal of Social Psychology*, *39*(2), 186–195.

Lander, K. (2008). Relating visual and vocal attractiveness for moving and static faces. *Animal Behaviour*, *75*(3), 817–822.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, *1*(2), 115–121.

Lawson, C., Lenz, G. S., Baker, A., & Myers, M. (2010). Looking like a winner: Candidate appearance and electoral success in new democracies. *World Politics*, *62*(4), 561–593.

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology*, *93*(2), 234–249.

Leary, T. (1957). *Interpersonal diagnosis of personality*. New York: Ronald Press.

Leder, H., Forster, M., & Gerger, G. (2011). The glasses stereotype revisited. *Swiss Journal of Psychology*, *70*(4), 211–222.

Lee, F., Hallahan, M., & Herzog, T. (1996). Explaining real-life events: How culture and domain shape attributions, *22*(7), 732–741.

Lefevre, C. E., Ewbank, M. P., Calder, A. J., von dem Hagen, E., & Perrett, D. I. (2013). It is all in the face: carotenoid skin coloration loses attractiveness outside the face. *Biology Letters*, *9*(6), 20130633.

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology*, *72*(2), 275–287.

Levy, B., & Langer, E. (1994). Aging free from negative stereotypes: Successful memory in China among the American deaf. *Journal of Personality and Social Psychology*, *66*(6), 989–997.

Lewicki, P. (1985). Nonconscious biasing effects of single instances on subsequent judgments. *Journal of Personality and Social Psychology*, *48*(3), 563–574.

Liao, H.-T. (2014). How much can one express in 140 characters? Comparison between English and other languages like Chinese. Retrieved from http://people.oii.ox.ac.uk/hanteng/2013/04/16/how-much-can-one-express-in-140-characters-comparison-between-english-and-other-languages-like-chinese/

Liermann, M., Steel, A., Rosing, M., & Guttorp, P. (2004). Random denominators and the analysis of ratio data. *Environmental and Ecological Statistics*, *11*(1), 55–71.

Light, L. L., Hollander, S., & Kayra-Stuart, F. (1981). Why attractive people are harder to remember. *Personality and Social Psychology Bulletin*, *7*(2), 269–276.

Li, J. (2002). A cultural model of learning Chinese "Heart and mind for wanting to learn." *Journal of Cross-Cultural Psychology*, *33*(3), 248–269.

Li, J. (2003). U.S and Chinese cultural beliefs about learning. *Journal of Educational Psychology*, *95*(2), 258–267.

Lin, L.-H., & Ho, Y.-L. (2009). Confucian dynamism, culture and ethical changes in Chinese societies – a comparative study of China, Taiwan, and Hong Kong. *The International Journal of Human Resource Management*, *20*(11), 2402–2417.

Lippmann, W. (1922). *Public Opinion*. Harcourt, Brace.

Little, A. C., Burriss, R. P., Jones, B. C., & Verosky, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior*, *28*(1), 18–27.

Little, A. C., Jones, B. C., & DeBruine, L. M. (2011). Facial attractiveness: evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*, 1638–1659.

Little, A. C., & Perrett, D. I. (2007). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology*, *98*(1), 111–126.

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology*, *82*(1), 5–18.

Löckenhoff, C. E., De Fruyt, F., Terracciano, A., McCrae, R. R., De Bolle, M., Costa, P. T., … Yik, M. (2009). Perceptions of Aging across 26 Cultures and their Culture-Level Associates. *Psychology and Aging*, *24*(4), 941–954.

Lutz, G. (2010). The electoral success of beauties and beasts. *Swiss Political Science Review*, *16*(3), 457–480.

Maass, A., Karasawa, M., Politi, F., & Suga, S. (2006). Do verbs and adjectives play different roles in different cultures? A cross-linguistic analysis of person representation. *Journal of Personality and Social Psychology*, *90*(5), 734–750.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, *51*(1), 93–120.

Maddox, K. B., & Gray Chase, S. (2004). Manipulating subcategory salience: exploring the link between skin tone and social perception of Blacks. *European Journal of Social Psychology*, *34*(5), 533–546.

Main, J. C., Jones, B. C., DeBruine, L. M., & Little, A. C. (2009). Integrating gaze direction and sexual dimorphism of face shape when perceiving the dominance of others. *Perception*, *38*(9), 1275–1283.

Mann, M. (1984). The autonomous power of the state: its origins, mechanisms and results. *European Journal of Sociology / Archives Européennes de Sociologie*, *25*(2), 185–213.

Marcella, G. (2004). Chapter 17: National security and the interagency process. In J. B. Bartholomees (Ed.), *US Army War College Guide to National Security Policy and Strategy*. Strategic Studies Institute.

Mark, L. S., Pittenger, J. B., Hines, H., Carello, C., Shaw, R. E., & Todd, J. T. (1980). Wrinkling and head shape as coordinated sources of age-level information. *Attention, Perception, & Psychophysics*, *27*(2), 117–124.

Masip, J., & Garrido, E. (2001). Is there a kernel of truth in judgements of deceptiveness? *Anales de Psicologia*, *17*(1), 101–120.

Mattavelli, G., Andrews, T. J., Asghar, A. U., Towler, J. R., & Young, A. W. (2012). Response of face-selective brain regions to trustworthiness and gender of faces. *Neuropsychologia*, *50*(9), 2205–2211.

Mattes, K., & Milazzo, C. (2014). Pretty faces, marginal races: Predicting election outcomes using trait assessments of British parliamentary candidates. *Electoral Studies*, *34*, 177–189.

MDBG dictionary. (2015). 慈祥. Retrieved January 30, 2015, from http://www.mdbg.net/chindict/chindict.php?page=worddict&wdrst=0&wdqb=%E6%85%88%E7%A5%A5.

Mehu, M., & Dunbar, R. I. (2008). Naturalistic observations of smiling and laughter in human group interactions. *Behaviour*, 1747–1780.

Mehu, M., Grammer, K., & Dunbar, R. I. (2007). Smiles when sharing. *Evolution and Human Behavior*, *28*(6), 415–422.

Mende-Siedlecki, P., Said, C. P., & Todorov, A. (2013). The social evaluation of faces: a meta-analysis of functional neuroimaging studies. *Social Cognitive and Affective Neuroscience*, *8*(3), 285–299.

Miller, J., & Hackley, S. A. (1992). Electrophysiological evidence for temporal overlap among contingent mental processes. *Journal of Experimental Psychology: General*, *121*(2), 195–209.

Ministry of Justice. (2014). *Statistics on women and the criminal justice system 2013*. Retrieved from https://www.gov.uk/government/statistics/women-and-the-criminal-justice-system-2013.

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.

Montepare, J. M., & Opeyo, A. (2002). The relative salience of physiognomic cues in differentiating faces: A methodological tool. *Journal of Nonverbal Behavior*, *26*(1), 43–59.

Montoya, R. M., & Horton, R. S. (2014). A two-dimensional model for the study of interpersonal attraction. *Personality and Social Psychology Review*, *18*(1), 59–86.

Morris, M. W., & Peng, K. (1994). Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality and Social Psychology*, *67*(6), 949–971.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479.

Mueller, U., & Mazur, A. (1996). Facial dominance of West Point cadets as a predictor of later military rank. *Social Forces*, *74*(3), 823–850.

Mukherjee, P., Whalley, H. C., McKirdy, J. W., Sprengelmeyer, R., Young, A. W., McIntosh, A. M., … Hall, J. (2014). Altered amygdala connectivity within the social brain in schizophrenia. *Schizophrenia Bulletin*, *40*(1), 152–160.

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin*, *35*(12), 1661–1671.

Nicol, K., Pope, M., Sprengelmeyer, R., Young, A. W., & Hall, J. (2013). Social judgement in borderline personality disorder. *PloS ONE*, *8*(11), e73440.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological Review*, *108*(2), 291–310.

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

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods*, *32*(3), 396–402.

Oda, R., Naganawa, T., Yamauchi, S., Yamagata, N., & Matsumoto-Oda, A. (2009). Altruists are trusted based on non-verbal cues. *Biology Letters*, *5*(6), 752–754.

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.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570.

Olivola, C. Y., & Todorov, A. (2010a). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*(2), 83–110.

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.

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

Osgood, C. E. (1964). Semantic differential technique in the comparative study of cultures. *American Anthropologist*, *66*(3), 171–200.

Osgood, C. E. (1969). On the whys and wherefores of E, P, and A. *Journal of Personality and Social Psychology*, *12*(3), 194–199.

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

Oyserman, D., Coon, H. M., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin*, *128*(1), 3–72.

Oyserman, D., Kemmelmeier, M., & Coon, H. M. (2002). Cultural psychology, A new look: reply to Bond (2002), Fiske (2002), Kitayama (2002), and Miller (2002). Retrieved from http://psycnet.apa.org/psycinfo/2002-00183-006.

Ozono, H., Watabe, M., Yoshikawa, S., Nakashima, S., Rule, N. O., Ambady, N., & Adams Jr, R. B. (2010). What's in a smile? Cultural differences in the effects of smiling on judgments of trustworthiness. *Letters on Evolutionary Behavioral Science*, *1*(1), 15–-18.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13.

Penton-Voak, I. S., & Chang, H. Y. (2008). Attractiveness judgements of individuals vary across emotional expression and movement conditions. *Journal of Evolutionary Psychology*, *6*(2), 89–100.

Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition*, *24*(5), 607–640.

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

Pittenger, J. B., & Shaw, R. E. (1975). Aging faces as viscal-elastic events: Implications for a theory of nonrigid shape perception. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(4), 374–382.

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/Revue Canadienne Des Sciences Du Comportement*, *40*(3), 171–177.

Porter, S., ten Brinke, L., & Gustaw, C. (2010). Dangerous decisions: the impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime & Law*, *16*(6), 477–491.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741–763.

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.

Quadflieg, S., & Macrae, C. N. (2011). Stereotypes and stereotyping: What's the brain got to do with it? *European Review of Social Psychology*, *22*(1), 215–273.

Quantcast. (2014). Linkedin.com's audience profile on Quantcast. Retrieved June 15, 2014, from https://www.quantcast.com/linkedin.com.

Quinn, K. A., & Macrae, C. N. (2011). The face and person perception: Insights from social cognition. *British Journal of Psychology*, *102*(4), 849–867.

Rahman, R. A., van Turennout, M., & Levelt, W. J. M. (2003). Phonological encoding is not contingent on semantic feature retrieval: An electrophysiological study on object naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 850–860.

Rapcsak, S. Z. (2013). Face recognition. In A. Chatterjee & H. B. Coslett (Eds.), *The Roots of Cognitive Neuroscience: Behavioral Neurology and Neuropsychology* (pp. 105–130). Oxford University Press.

Ratner, K. G., Dotsch, R., Wigboldus, D. H. J., van Knippenberg, A., & Amodio, D. M. (2014). Visualizing minimal ingroup and outgroup faces: Implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology*, *106*(6), 897–911.

Read, S. J., & Miller, L. C. (1998). On the dynamic construction of meaning: An interactive activation and competition model of social perception. *Connectionist Models of Social Reasoning and Social Behavior*, 27–68.

Reed, L. I., Zeglen, K. N., & Schmidt, K. L. (2012). Facial expressions as honest signals of cooperative intent in a one-shot anonymous Prisoner's Dilemma game. *Evolution and Human Behavior*, *33*(3), 200–209.

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS ONE*, *7*(3), e34293.

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, *57*(1), 199–226.

Rhodes, G., Brake, S., Taylor, K., & Tan, S. (1989). Expertise and configural coding in face recognition. *British Journal of Psychology*, *80*(3), 313–331.

Rhodes, G., Lie, H. C., Thevaraja, N., Taylor, L., Iredell, N., Curran, C., … Simmons, L. W. (2011). Facial attractiveness ratings from video-clips and static images tell the same story. *PLoS ONE*, *6*(11), e26653.

Riggio, H. R., & Riggio, R. E. (2010). Appearance-based trait inferences and voting: Evolutionary roots and implications for leadership. *Journal of Nonverbal Behavior*, *34*(2), 119–125.

Roberts, S. C., Saxton, T. K., Murray, A. K., Burriss, R. P., Rowland, H. M., & Little, A. C. (2009). Static and dynamic facial images cue similar attractiveness judgements. *Ethology*, *115*(6), 588–595.

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, *9*(4), 283–294.

Rossion, B., & Michel, C. (2011). An experience-based holistic account of the other-race face effect. In A. J. Calder, G. Rhodes, & M. Johnson (Eds.), *Oxford Handbook of Face Perception* (pp. 215–244). Oxford: Oxford University Press.

Rothbart, M. (1988). Categorisation and impression formation: Capturing the mind's flexibility. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Advances in Social Cognition* (Vol. 1, pp. 139–144). Hillsdale, NJ: Erlbaum.

Rubenstein, A. J. (2005). Variation in perceived attractiveness differences between dynamic and static faces. *Psychological Science*, *16*(10), 759–762.

Rudman, L. A. (1998). Self-promotion as a risk factor for women: the costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology*, *74*(3), 629–645.

Rudman, L. A., & Glick, P. (1999). Feminized management and backlash toward agentic women: the hidden costs to women of a kinder, gentler image of middle managers. *Journal of Personality and Social Psychology*, *77*(5), 1004–1010.

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

Rudman, L. A., Moss-Racusin, C. A., Glick, P., & Phelan, J. E. (2012). Reactions to vanguards: Advances in backlash theory. *Advances in Experimental Social Psychology*, *45*, 167–227.

Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E., & Nauts, S. (2012). Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *Journal of Experimental Social Psychology*, *48*(1), 165–179.

Rudoy, J. D., & Paller, K. A. (2009). Who can you trust? Behavioral and neural differences between perceptual and memory-based influences. *Frontiers in Human Neuroscience*, *3*(16), 1–6.

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.

Rule, N. O., & Ambady, N. (2009). She's got the look: Inferences from female chief executive officers' faces predict their success. *Sex Roles*, *61*(9-10), 644–652.

Rule, N. O., & Ambady, N. (2010). Democrats and Republicans can be differentiated from their faces. *PLoS ONE*, *5*(1), e8733.

Rule, N. O., Ambady, N., Adams Jr, R. B., Ozono, H., Nakashima, S., Yoshikawa, S., & Watabe, M. (2010). Polling the face: Prediction and consensus across cultures. *Journal of Personality and Social Psychology*, *98*(1), 1–15.

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 of Experimental Social Psychology*, *45*(6), 1245–1251.

Rule, N. O., Garrett, J. V., Ambady, N., & Rustichini, A. (2010). On the perception of religious group membership from faces. *PLoS ONE*, *5*(12), e14241.

Rule, N. O., Ishii, K., & Ambady, N. (2011). Cross-cultural impressions of leaders' faces: Consensus and predictive validity. *International Journal of Intercultural Relations*, *35*, 833–841.

Rule, N. O., Ishii, K., Ambady, N., Rosen, K. S., & Hallett, K. C. (2011). Found in translation: Cross-cultural consensus in the accurate categorization of male sexual orientation. *Personality and Social Psychology Bulletin*, *37*(11), 1499–1507.

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology*, *104*(3), 409–426.

Rule, N. O., & Tskhay, K. O. (2014). The influence of economic context on the relationship between chief executive officer facial appearance and company profits. *The Leadership Quarterly*, *25*(5), 846–854.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Russell, R. (2009). A sex difference in facial contrast and its exaggeration by cosmetics. *Perception*, *38*(8), 1211–1219.

Russell, R. (2010). Why cosmetics work. *The Science of Social Vision*, *1*(9), 186–204.

Ryan, E. B., Jin, Y., Anas, A. P., & Luh, J. J. (2004). Communication beliefs about youth and old age in Asia and Canada. *Journal of Cross-Cultural Gerontology*, *19*(4), 343–360.

Sadalla, E. K., Kenrick, D. T., & Vershure, B. (1987). Dominance and heterosexual attraction. *Journal of Personality and Social Psychology*, *52*(4), 730–738.

Said, C. P., Dotsch, R., & Todorov, A. (2010). The amygdala and FFA track both social and non-social face dimensions. *Neuropsychologia*, *48*, 3596–3605.

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, *9*(2), 260–264.

Santos, I. M., & Young, A. W. (2005). Exploring the perception of social characteristics in faces using the isolation effect. *Visual Cognition*, *12*(1), 213–247.

Santos, I. M., & Young, A. W. (2008). Effects of inversion and negation on social inferences from faces. *Perception*, *37*(7), 1061–1078.

Santos, I. M., & Young, A. W. (2011). Inferring social attributes from different face regions: Evidence for holistic processing. *The Quarterly Journal of Experimental Psychology*, *64*(4), 751–766.

Saucier, G., Thalmayer, A. G., & Bel-Bahar, T. S. (2014). Human attribute concepts: Relative ubiquity across twelve mutually isolated languages. *Journal of Personality and Social Psychology*, *107*(1), 199–216.

Schneider, D. J. (2004). *The psychology of stereotyping*. NY: Guildford Press.

Sczesny, S., & Kühnen, U. (2004). Meta-cognition about biological sex and gender-stereotypic physical appearance: Consequences for the assessment of leadership competence. *Personality and Social Psychology Bulletin*, *30*(1), 13–21.

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.

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., & Bevan, W. (1956). Personalities in faces: III. A cross-cultural comparison of impressions of physiognomy and personality in faces. *The Journal of Social Psychology*, *43*(2), 283–288.

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.

Sedghi, A. (2014). Facebook: 10 years of social networking, in numbers. Retrieved June 15, 2014, from http://www.theguardian.com/news/datablog/2014/feb/04/facebook-in-numbers-statistics.

Sell, A., Cosmides, L., Tooby, J., Sznycer, D., von Rueden, C., & Gurven, M. (2009). Human adaptations for the visual assessment of strength and fighting ability from the body and face. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1656), 575–584.

Shaffer, D. R., Crepaz, N., & Sun, C.-R. (2000). Physical attractiveness stereotyping in cross-cultural perspective similarities and differences between Americans and Taiwanese. *Journal of Cross-Cultural Psychology*, *31*(5), 557–582.

Six, B., & Eckes, T. (1991). A closer look at the complex structure of gender stereotypes. *Sex Roles*, *24*(1), 57–71.

Spears, R., & Haslam, S. A. (1997). Stereotyping and the burden of cognitive load. In R. Spears, P. J. Oakes, N. Ellemers, & S. A. Haslam (Eds.), *The Social Psychology of Stereotyping and Group Life* (pp. 171–207). Wiley-Blackwell.

Stangor, C., Lynch, L., Duan, C., & Glas, B. (1992). Categorization of individuals on the basis of multiple social features. *Journal of Personality and Social Psychology*, *62*(2), 207–218.

Stephen, I. D., Law Smith, M. J., Stirrat, M. R., & Perrett, D. I. (2009). Facial skin coloration affects perceived health of human faces. *International Journal of Primatology*, *30*(6), 845–857.

Stephen, I. D., & McKeegan, A. M. (2010). Lip colour affects perceived sex typicality and attractiveness of human faces. *Perception*, *39*(8), 1104–1110.

Stillman, T. F., Maner, J. K., & Baumeister, R. F. (2010). A thin slice of violence: Distinguishing violent from nonviolent sex offenders at a glance. *Evolution and Human Behavior*, *31*(4), 298–303.

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*(3), 349–354.

Strom, M. A., Zebrowitz, L. A., Zhang, S., Bronstad, P. M., & Lee, H. K. (unpublished manuscript). Race-related facial qualities contribute to stereotyping by White, Black, and Korean Judges. *Brandeis University*. Retrieved from http://www.cogsci.ucsd.edu/~s6zhang/publications/face.pdf.

Strom, M. A., Zebrowitz, L. A., Zhang, S., Bronstad, P. M., & Lee, H. K. (2012). Skin and bones: The contribution of skin tone and facial structure to racial prototypicality ratings. *PLoS ONE*, *7*(7), e41193.

Sutherland, C. A. M. (2015). A basic guide to Psychomorph. University of York. Retrieved from https://dlib.york.ac.uk/yodl/app/home/detail?id=york:847354&ref=browse.

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118.

Sutherland, C. A. M., Oldmeadow, J. A., & Young, A. W. (in prep). Convergent and divergent impressions of male and female faces.

Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (in press). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology*.

Thagard, P., & Kunda, Z. (1998). Making sense of people: Coherence mechanisms. In S. J. Read & L. C. Miller (Eds.), *Connectionist Models of Social Reasoning and Social Behavior*. New Jersey: Lawrence Erlbaum Associates.

Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in Cognitive Sciences*, *3*(12), 452–460.

Thornton, G. R. (1944). The effect of wearing glasses upon judgments of personality traits of persons seen briefly. *Journal of Applied Psychology*, *28*(3), 203–207.

Tiddeman, B., Burt, M., & Perrett, D. I. (2001). Prototyping and transforming facial textures for perception research. *Computer Graphics and Applications, IEEE*, *21*(5), 42–50.

Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences*, *1124*(1), 208–224.

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Social Cognitive and Affective Neuroscience*, *3*(2), 119–127.

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.

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.

Todorov, A., & Duchaine, B. (2008). Reading trustworthiness in faces without recognizing faces. *Cognitive Neuropsychology*, *25*(3), 395–410.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626.

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*(1), 519–545.

Todorov, A., & Oosterhof, N. N. (2011). Modeling social perception of faces. *IEEE Signal Processing Magazine*, *28*(2), 117–122.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*(6), 813–833.

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science*, *25*(7), 1404–1417.

Todorov, A., Said, C., Engell, A., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460.

Todorov, A., Said, C. P., & Verosky, S. C. (2011). Personality impressions from facial appearance. In A. Calder, G. Rhodes, M. Johnson, & J. V. Haxby (Eds.), *Oxford Handbook of Face Perception.* (pp. 630–651). Oxford: Oxford University Press.

Tracy, J. L., & Randles, D. (2011). Four models of basic emotions: A review of Ekamn and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review, 3(4),* 397-405.

Triandis, H. C. (1995). *Individualism & collectivism.* Westview Press.

Vakhrusheva, J., Zemon, V., Bar, M., Weiskopf, N. G., Tremeau, F., Petkova, E., … Butler, P. D. (2014). Forming first impressions of others in schizophrenia: Impairments in fast processing and in use of spatial frequency information. *Schizophrenia Research*, *160*(1), 142–149.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321–327.

Verhulst, B., Lodge, M., & Lavine, H. (2010). The attractiveness halo: Why some candidates are perceived more favorably than others. *Journal of Nonverbal Behavior*, *34*(2), 111–117.

Vernon, R. J. W., Sutherland, C. A. M., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *PNAS*, *111*(32), E3353–E3361.

Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science*, *21*(6), 779–785.

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.

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.

Vigil, J. M. (2009). A socio-relational framework of sex differences in the expression of emotion. *Behavioral and Brain Sciences*, *32*(5), 375–390.

Vonk, R. (1994). Trait inferences, impression formation, and person memory: Strategies in processing inconsistent information about persons. *European Review of Social Psychology*, *5*(1), 111–149.

Vuilleumier, P., & Sander, D. (2008). Trust and valence processing in the amygdala. *Social Cognitive and Affective Neuroscience*, *3*(4), 299–302.

Walker, M., Jiang, F., Vetter, T., & Sczesny, S. (2011). Universals and cultural differences in forming personality trait judgments from faces. *Social Psychological and Personality Science*, *2*(6), 609–617.

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.

Walker Naylor, R. (2007). Nonverbal cues-based first impressions: Impression formation through exposure to static images. *Marketing Letters*, *18*(3), 165–179.

Wall, H. J., Taylor, P. J., Dixon, J., Conchie, S. M., & Ellis, D. A. (2013). Rich contexts do not always enrich the accuracy of personality judgments. *Journal of Experimental Social Psychology*, *49*(6), 1190–1195.

Warner, R. M. (2013). *Applied Statistics: From Bivariate Through Multivariate Techniques* (Second). Thousand Oaks, CA: Sage Publications.

Watkins, C. D., Jones, B. C., & DeBruine, L. M. (2010). Individual differences in dominance perception: Dominant men are less sensitive to facial cues of male dominance. *Personality and Individual Differences*, *49*(5), 967–971.

Wheeler, L., & Kim, Y. (1997). What is beautiful is culturally good: The physical attractiveness stereotype has different content in collectivistic cultures. *Personality and Social Psychology Bulletin*, *23*(8), 795–800.

White, C., & Edgar, G. (2010). Inequalities in healthy life expectancy by social class and area type: England, 2001-03. *Health Statistics Quarterly / Office for National Statistics*, (45), 28–56.

White, M. J., Kruczek, T. A., Brown, M. T., & White, G. B. (1989). Occupational sex stereotypes among college students. *Journal of Vocational Behavior*, *34*(3), 289–298.

White, M. J., & White, G. B. (2006). Implicit and explicit occupational gender stereotypes. *Sex Roles*, *55*(3-4), 259–266.

White, S., Hill, E., Winston, J., & Frith, U. (2006). An islet of social ability in Asperger Syndrome: judging social attributes from faces. *Brain and Cognition*, *61*(1), 69–77.

Whitley Jr, B. E. (1999). Right-wing authoritarianism, social dominance orientation, and prejudice. *Journal of Personality and Social Psychology*, *77*(1), 126–134.

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, *37*(3), 395–412.

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.

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology*, *67*(2), 222–232.

Wojciszke, B. (2005). Morality and competence in person-and self-perception. *European Review of Social Psychology*, *16*(1), 155–188.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263.

Wolffhechel, K., Fagertun, J., Jacobsen, U. P., Majewski, W., Hemmingsen, A. S., Larsen, C. L., … Jarmer, H. (2014). Interpretation of appearance: The effect of facial features on first impressions and personality. *PLoS ONE*, *9*(9), e107721.

World Population Clock: 7 Billion People (2014) - Worldometers. (2014). Retrieved July 23, 2014, from http://www.worldometers.info/world-population/#region.

Xu, F., Wu, D., Toriyama, R., Ma, F., Itakura, S., & Lee, K. (2012). Similarities and differences in Chinese and Caucasian adults' use of facial cues for trustworthiness judgments. *PloS One*, *7*(4), e34859.

Yang, D., Qi, S., Ding, C., & Song, Y. (2011). An ERP study on the time course of facial trustworthiness appraisal. *Neuroscience Letters*, *496*(3), 147–151.

Yang, S.-Y., & Sternberg, R. J. (1997a). Conceptions of intelligence in ancient Chinese philosophy. *Journal of Theoretical and Philosophical Psychology*, *17*(2), 101–119.

Yang, S.-Y., & Sternberg, R. J. (1997b). Taiwanese Chinese people's conceptions of intelligence. *Intelligence*, *25*(1), 21–36.

Yang, X. (2014). The role of photographs in online peer-to-peer lending behavior. *Social Behavior and Personality*, *42*(3), 445–452.

Ybarra, O., Chan, E., Park, H., Burnstein, E., Monin, B., & Stanik, C. (2008). Life's recurring challenges and the fundamental dimensions: An integration and its implications for cultural differences and similarities. *European Journal of Social Psychology*, *38*(7), 1083–1092.

Yoon, C., Hasher, L., Feinberg, F., Rahhal, T. A., & Winocur, G. (2000). Cross-cultural differences in memory: The role of culture-based stereotypes about aging. *Psychology and Aging*, *15*(4), 694–704.

YouGov. (2014). Online Dating Services. Retrieved June 11, 2014, from http://yougov.co.uk/news/2014/02/13/seven-ten-online-dating-virgins-willing-try-findin/.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., & Perrett, D. I. (1997). Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition*, *63*(3), 271–313.

Zebrowitz, L. A. (2005). Appearance DOES Matter. *Science*, *308*, 1565–1566.

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.

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.

Zebrowitz, L. A., & Franklin Jr, R. G. (2014). The attractiveness halo effect and the babyface stereotype in older and younger adults: Similarities, own-age accentuation, and older adult positivity effects. *Experimental Aging Research*, *40*(3), 375–393.

Zebrowitz, L. A., Franklin Jr, R. G., Hillman, S., & Boc, H. (2013). Older and younger adults' first impressions from faces: Similar in agreement but different in positivity. *Psychology and Aging*, *28*(1), 202–212.

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.

Zebrowitz, L. A., Kikuchi, M., & Fellous, J. M. (2007). Are effects of emotion expression on trait impressions mediated by babyfaceness? Evidence from connectionist modeling. *Personality and Social Psychology Bulletin*, *33*(5), 648–662.

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.

Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior*, *15*(6), 603–623.

Zebrowitz, L. A., & Montepare, J. M. (1992). Impressions of babyfaced individuals across the life span. *Developmental Psychology*, *28*(6), 1143–1152.

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, *2*(3), 1497–1517.

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.

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.

Zebrowitz, L. A., Voinescu, L., & Collins, M. A. (1996). “Wide-eyed” and “crooked-faced”: Determinants of perceived and real honesty across the life span. *Personality and Social Psychology Bulletin*, *22*(12), 1258–1269.

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*, 486–496.

Zebrowitz, L. A., Wang, R., Bronstad, P. M., Eisenberg, D., Undurraga, E., Reyes-García, V., & Godoy, R. (2012). First impressions from faces among US and culturally isolated Tsimane’people in the Bolivian rainforest. *Journal of Cross-Cultural Psychology*, *43*(1), 119–134.