Cognition 127 (2013) 105–118

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Social inferences from faces: Ambient images generate a three-dimensional model

Clare A.M. Sutherland [a,*], Julian A. Oldmeadow [a], Isabel M. Santos [b], John Towler [c], D. Michael Burt [d], Andrew W. Young [a]

[a] Department of Psychology, University of York, Heslington, York YO10 5DD, UK
[b] Departamento de Educação, Universidade de Aveiro, Campus Universitário de Santiago, Portugal
[c] Department of Psychological Sciences, Birkbeck College, University of London, UK
[d] Department of Psychology, University of Durham, UK

ARTICLE INFO

Article history:
Received 31 July 2012
Revised 29 November 2012
Accepted 4 December 2012

Keywords:
Social inferences
Face perception
First impressions

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, White, Van Montfort, & Burton, 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.

1. Introduction

1.1. 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, Mandisodza, Goren, & Hall, 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, Kikuchi, & Fellous, 2010). The dominance

* Corresponding author. Tel.: +44 1904 322861.
E-mail address: cs770@york.ac.uk (C. Sutherland).

0010-0277/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2012.12.001

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, Neave, & Seydel, 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 and 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, Fiske, & Glick, 2008; see also 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, Jones, & DeBruine, 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, Berscheid, & Walster, 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, Ashmore, Makhijani, & Longo, 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.

1.2. 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, Jenkins, & Schweinberger, 2011). One way to maintain such variability is to sample publically available, pre-existing photographs from the internet. Jenkins et al. (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 exam-

ple, 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.

1.3. 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, and Todorov (2010) for a direct manipulation of face shape along the valence dimension).

2. 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; Tiddeman, Burt, and 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.

2.1. Method 1

2.1.1. 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, 2008, 2011). The photographs in this database are standardised to be 150 pixels in height (approximately 5 cm 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, Young, Sacco, & Bernstein, 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 750 ms. Participants were given as much time as they wanted but were encouraged to go with their 'gut instinct' (Todorov et al., 2005).

2.1.2. Participants

Twelve 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.

2.1.3. Stimuli and 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 Fig. 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).

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 Fig. 1). These continua allowed us to assess, using correlations, whether

Fig. 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%.

the averages were reliably perceived as changing on their respective manipulated traits.

2.1.4. 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 carry-over 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.

2.2. Results 1

2.2.1. 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.

2.2.2. 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; 11 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. 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 (Fig. 1). Indeed, while the averages are not as controlled as those from previous research using more homogenous initial stimulus sets (e.g. Penton-Voak, Pound, Little, & Perrett, 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, 2012). 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 (c.f. 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.

3. 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 Section 2.1.1 initial ratings collection) so that enough variables could potentially load on each factor to make them meaningful (Kline, 1994). As explained previously (Section 2.1.1), 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.

3.1. 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.

3.2. Results 2

3.2.1. 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 S1, in the Supplementary material, for the correlational matrix).

3.2.2. 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 1; for further information see Table S2, Supplementary material).

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.

Table 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.

3.2.3. Model robustness

To ascertain the model robustness, different analyses were implemented and various traits excluded (see Table S3, 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.

3.2.4. 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 S3) and all confirmatory analyses were carried out on the other half. Multivariate normality was acceptable (Byrne, 2009) and the maximum likelihood method was employed (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

Table 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

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 chi^2, 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). 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 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 (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.

3.3. 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.

4. 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).

4.1. Methods 3

4.1.1. Participants

Thirty 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.

4.1.2. 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 (Fig. 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.

To map the model space more fully, new averages were then generated by averaging together (a) all possible pairs

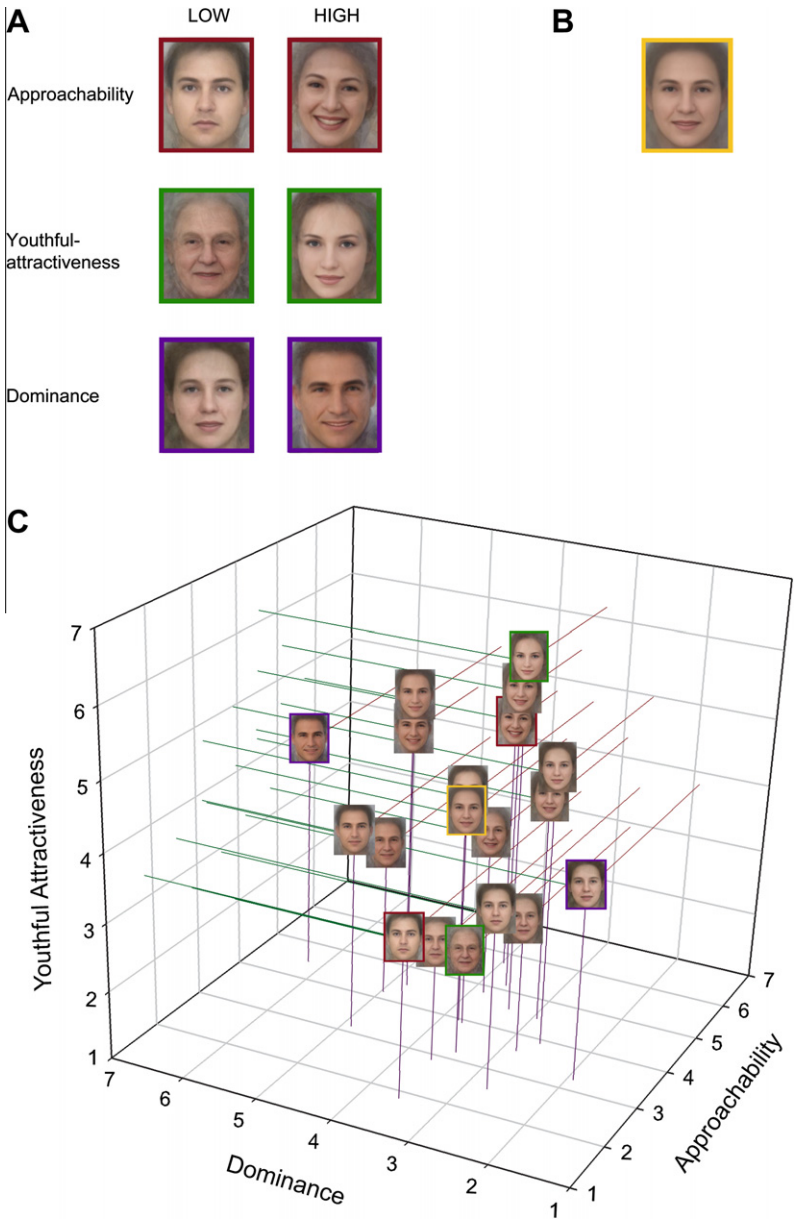

Fig. 2. Experiment 3 stimuli: (A) High and low face averages for each of the three dimensions. Each is an average of the 20 highest or lowest factor-scoring face photographs on a dimension. (B) The grand-origin face average, which is an average composed of the original 6 face averages. (C) 19 out of 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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 (Fig. 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 Fig. 2C.

4.1.3. 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 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.

4.1.4. 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.

4.2. Results 3

4.2.1. 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.

4.2.2. 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 (Fig. 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 $n = 20$, $p < .001$).

4.3. 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 (Fig. 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

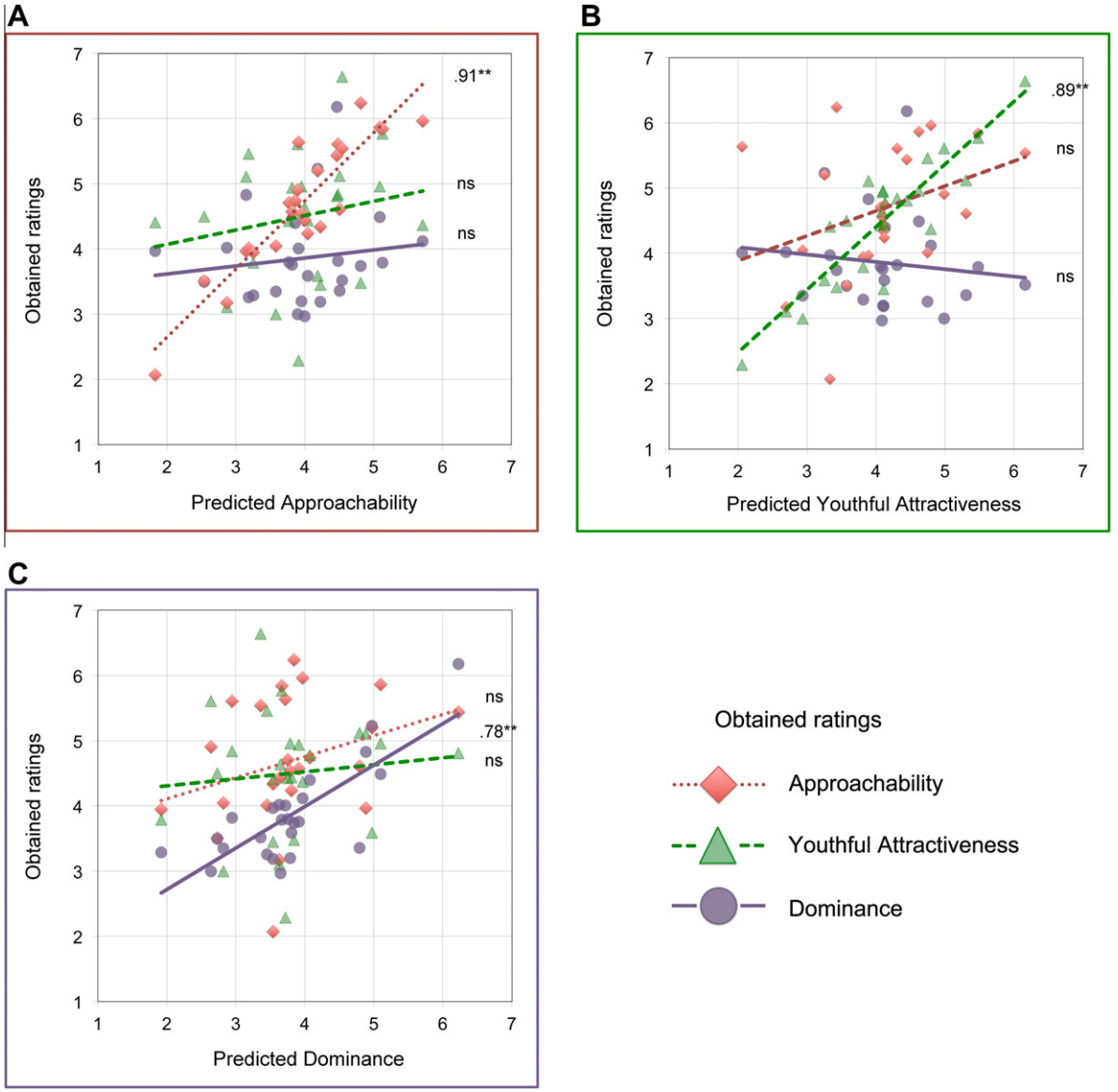

Fig. 3. Correlations between the three sets of (averaged) obtained ratings separately with predicted approachability (A), youthful-attractiveness (B) and dominance (C). Each point represents a face average. $^{**}p < .001$.

low approachability average appears to be male, and either neutral or negatively valenced. This corresponds with the current analyses and with previous research (Hess, Adams, & Kleck, 2004; Said, Sebe, & Todorov, 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 (Fig. 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.

5. 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 and 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, Grammer, & Thornhill, 2001) and these 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., 1999, 2000) 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 et al. (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.

5.1. 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.

Acknowledgements

We are grateful to the journal's Action Editor and three reviewers for insightful comments on a previous draft. Portions of this paper were based on the first author's master's thesis at the Department of Psychology, York, UK. The research was funded by an ESRC studentship to the first author. The funding source had no influence on the research.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2012.12.001.

References

American Society of Plastic Surgeons (2010). *Report of the 2010 plastic surgery statistics*. <http://www.plasticsurgery.org/News-and-Resources/Statistics.html> Retrieved May 2011.

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

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.

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

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).

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.

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

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

Bruce, V., & Young, A. (2012). *Face perception*. Psychology Press.

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

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.

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.

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. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science, 3*(5), 562–571.

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.

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.

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.

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.

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.

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

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

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

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

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.

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

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.

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*. Oxford: Oxford University.

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

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

Krings, F., Sczesny, Sabine, & Kluge, A. (2011). Stereotypical inferences as mediators of age discrimination: The role of competence and warmth. *British Journal of Management, 22*(2), 187–201.

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.

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.

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.

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., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

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.

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology, 57*(1), 199–226.

Rossion, B., & Michel, C. (2011). An experience-based account of the other-race face effect. In A. J. Calder, G. Rhodes, & M. Johnson (Eds.), *Oxford handbook of face perception*. Oxford: Oxford University.

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.

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.

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., 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., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623–1626.

Todorov, A., & Oosterhof, N. N. (2011). Modeling social perception of faces. *IEEE Signal Processing Magazine, 28*(2), 117–122.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321–327.

Vigil, J. M. (2009). A socio-relational framework of sex differences in the expression of emotion. *Behavioral and Brain Sciences, 32*(05), 375–390.

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.

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.

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology, 37*(3), 395–412.

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.

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).

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., & Montepare, J. M. (1992). Impressions of babyfaced individuals across the life span. *Developmental Psychology, 28*(6), 1143–1152.

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.