

Internal facial features are signals of personality and health

Robin S. S. Kramer and Robert Ward

Bangor University, Bangor, UK

We investigated forms of socially relevant information signalled from static images of the face. We created composite images from women scoring high and low values on personality and health dimensions and measured the accuracy of raters in discriminating high from low trait values. We also looked specifically at the information content within the internal facial features, by presenting the composite images with an occluding mask. Four of the Big Five traits were accurately discriminated on the basis of the internal facial features alone (conscientiousness was the exception), as was physical health. The addition of external features in the full-face images led to improved detection for extraversion and physical health and poorer performance on intellect/imagination (or openness). Visual appearance based on internal facial features alone can therefore accurately predict behavioural biases in the form of personality, as well as levels of physical health.

Keywords: Face; Evolution; Personality; Health; Signal.

The face can be used to predict a person's behaviour. Some transient emotional states, such as surprise or fear, are indicated by motion within the face. And although we are frequently warned that appearances are deceiving, recent evidence also suggests that static properties of the face are similarly expressive. Some reviewers have suggested that the face is a visible indicator of sex hormone levels (e.g., Fink & Penton-Voak, 2002; Johnston, Hagel, Franklin, Fink, & Grammer, 2001). To the extent that sex hormones direct action, the face would then be a predictor of any such hormonally driven behaviours. For example, Swaddle and Reiersen (2002) note that levels of testosterone in men are associated with both the development of the jaw-line and levels of aggressive behaviour

(Mazur & Booth, 1998). Shape of the jaw may therefore be an accurate predictor of dominance behaviours in men (Swaddle & Reiersen, 2002). In women, ovulation is associated with both visible changes in facial attractiveness (Penton-Voak et al., 1999) and a change in sexual interests and potential sexual behaviours (Gangestad, Thornhill, & Garver, 2002).

More recently, static properties of the face have been associated with enduring behavioural biases in the form of personality. Research has found that raters could identify certain personality traits of strangers (individually or in the form of composites) at a level significantly above chance, based only on a photograph of the face with a neutral expression (Little & Perrett, 2007; Penton-Voak,

Correspondence should be addressed to Robin S. S. Kramer, School of Psychology, Bangor University, Bangor, Gwynedd, UK.
E-mail: psp837@bangor.ac.uk

Pound, Little, & Perrett, 2006; Shevlin, Walker, Davies, Banyard, & Lewis, 2003). Using composites based on the Big Five traits, extraversion, conscientiousness, and agreeableness were identified accurately (Little & Perrett, 2007). Boothroyd, Jones, Burt, DeBruine, and Perrett (2008) have shown that indications of sociosexual orientation are also available from static face images. So not only are people quite willing to make personality and other judgements on the basis of appearance and other “thin slices”, but these judgements can be accurate.

The face may also provide a visible signal of health, although the picture is not yet certain. A powerful theoretical standpoint is that preferences in attractiveness have evolved to guide mate choice. By this perspective, attractiveness should be a useful cue to traits of great adaptive importance, such as fertility and health (e.g., Grammer, Fink, Møller, & Thornhill, 2003). Significant effects of facial attractiveness and health have been found (e.g., Rhodes, Chan, Zebrowitz, & Simmons, 2003), but also a concealing effect has been reported, in that ratings of health are more accurate when effects of attractiveness are partialled out (Kalick, Zebrowitz, Langlois, & Johnson, 1998). Further examination of the database used by Kalick et al. suggested a small correlation between health and averageness (an r around $-.1$ between health and “distinctiveness”), but no relationship with face symmetry (Rhodes et al., 2001). There was no correlation of perceived femininity with actual health, but a small correlation between perceived masculinity and health (Rhodes et al., 2003). An ongoing debate is therefore the extent to which facial attractiveness indicates health—for example, there is disagreement about the importance of fluctuating asymmetry as an indicator of health (see Weeden & Sabini, 2005, and a response from Grammer, Fink, Møller, & Manning, 2005).

With these findings in mind, we decided to investigate whether faces accurately signal health and personality. Given the importance of their findings, our first aim was to replicate the main results of Penton-Voak et al. (2006) and Little and Perrett (2007), showing that aspects of

personality were discernible from static facial images alone. Specifically, we looked to see whether composite images, formed from women with high and low personality trait values, could be accurately identified. For example, when presented with one composite made from the faces of extraverted women and another composite made from introverted women, could observers identify which is which? Second, we wished to look again at the issue of health and appearance. Here we were specifically interested in the relationship between attractiveness and health (e.g., Grammer et al., 2003), but in some sense the more fundamental issue of whether health can be accurately estimated from the face. We therefore looked to see whether composite face images from women of high and low self-reported health could be accurately identified. The use of composite images would effectively minimize any influence of fluctuating asymmetry, so that accurate health identification would have to rest on other factors.

Finally, we sought to develop a method for determining where in the face information relating to personality and/or health was carried. Specifically, we tested (a) whether the internal features corresponding to the area around the eyes, nose, and mouth were sufficient for trait recognition; and (b) for which traits did other, noninternal, features contribute to identification?

GENERAL METHOD

Experiments 1 and 2 each consisted of a short rating task of about 5 to 10 minutes. Participants completed both experiments, presented in counterbalanced order between participants. For exposition, it is simpler and clearer to consider the results of each task as separate experiments.

Stimuli: The composite images

For both experiments, composite face images were made from facial photos and inventories of personality and health, taken from a pool of 63 Caucasian women undergraduates (age $M = 21.03$, $SD =$

1.94, age unavailable for 4 participants). Course credit was given for participation. Each woman completed the Mini-IPIP (IPIP: International Personality Item Pool) personality inventory (Donnellan, Oswald, Baird, & Lucas, 2006) and the Short-Form 12-Item Health Survey (SF-12; Ware, Kosinski, & Keller, 1996). The SF-12 provides both a physical component summary (PCS) and a mental component summary (MCS). Digital photographs of each woman's face were taken by a professional photographer using professional-quality camera, lighting, and reflectors. Photos were constrained to reflect neutral expression, eyes on the camera; consistent posture, lighting, and distance to the camera; no glasses; jewellery, or make-up if possible; and hair back.

The 15 highest and lowest scorers were identified on each of seven traits: Big Five traits from the Mini-IPIP (agreeableness, extraversion, conscientiousness, neuroticism, and intellect/imagination) and physical and mental health (based upon the PCS and MCS subscales of the SF-12). Separate composite images were made for the high and low scorers using Abrosoft FantaFace Mixer, based on 112 key locations within the face and around the face outline. In addition, an average composite face was made for the entire group of 63 women. (All composite images can be seen in Figure 1.)

Differences between traits

Not surprisingly, the participants selected for the high and low composites differed significantly along the selected trait (all p s < .001), but there were also a few other differences. As might be expected, there was overlap in the measures of mental health, as assessed by the SF-12, and neuroticism. The high mental health group had significantly lower neuroticism than the low mental health group, $t(28) = 5.00, p < .001$; and likewise, the low neuroticism group had higher mental health scores than did the high neuroticism group, $t(28) = 6.77, p < .001$. The overlap of these measures simply reflects their similar domains. In addition, the high extraversion group had significantly higher mental health scores than the low extraversion group, $t(28) = 2.44,$

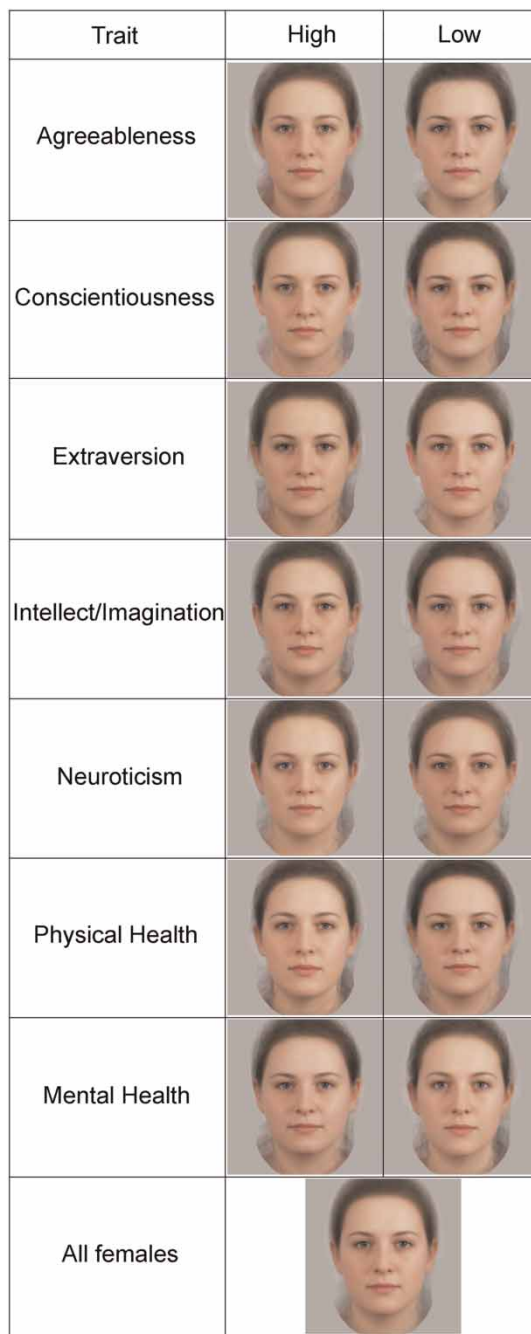


Figure 1. Composite faces based on self-reported personality (Mini-IPIP, International Personality Item Pool; Donnellan et al., 2006) and health (SF-12, Short-Form 12-Item Health Survey; Ware et al., 1996). To view a colour version of this figure, please see the online issue of the Journal.

$p = .021$; and the low agreeableness group had lower conscientiousness scores than the high agreeableness group, $t(28) = 2.17$, $p = .038$. The potential implications of these differences are considered later. There were no other significant differences.

Internal face images

The composites were converted to greyscale to minimize any skin tone differences and were cropped to produce images where only the internal features were visible (see Figure 2). By presenting only this limited region of the composites, we could explore whether the internal features of the face alone carried both health and personality information.

EXPERIMENT 1: ACCURACY OF TRAIT IDENTIFICATION FROM FULL AND INTERNAL FACES

Here we measured accuracy in discriminating composites made from the high and low value scorers on each trait. The high and low value composite faces were presented together, along with a discrimination question relevant to the trait (see Figure 3). Participants judged which of the two faces better fitted the question.

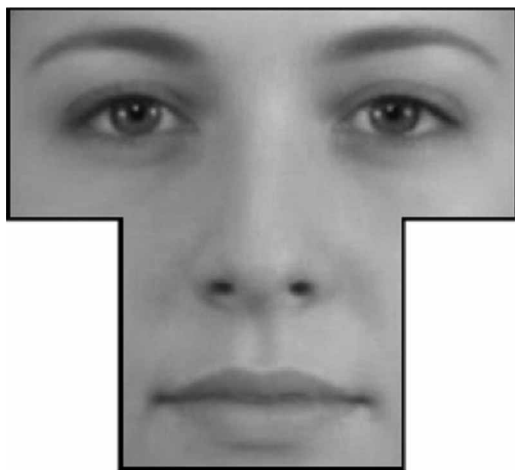


Figure 2. The “high physical health” composite, converted to black and white and cropped so that only the internal features are visible.

Method

Design

The experiment was defined by two factors describing the stimulus images: Trait (agreeableness, conscientiousness, extraversion, neuroticism, intellect/imagination, physical health, mental health) \times Face Type (full face or internal features only). Face type was varied between participants, trait within.

Participants

There were 131 participants (92 females; age $M = 20.99$, $SD = 2.33$), including 59 of the 63 women who contributed to the stimulus creation pool. These women plus 31 men completed the experiment using the full faces, for class credit. The remaining 41 participants (33 females) were not in the class and completed the task using the internal face composites for printing credits. All participants were undergraduate students in the Psychology programme at Bangor University.

Procedure

On each of the 28 trials, the high and low composites for a trait were presented to the participant (image size of 489×489 pixels, or about 13×13 cm on a 96-dpi screen), one to the left and one to the right of centre. Viewing distance was not fixed. The task was to judge which face better suited the discrimination statement appearing beneath the composite pair. Participants indicated their answer using the mouse to click on the appropriate image, and the next trial then appeared. The experiment was self-paced, and participants were encouraged to make their best answer.

Each composite pair was presented four times, each time with a different discrimination statement. For the Big Five traits, the discrimination statements were taken directly from the four relevant questions of the Mini-IPIP inventory used for scoring the women in the stimulus pool. For physical and mental health, the discrimination statements were taken from four items of the PCS and MCS subscales of the SF-12. The four items chosen were the ones producing the largest

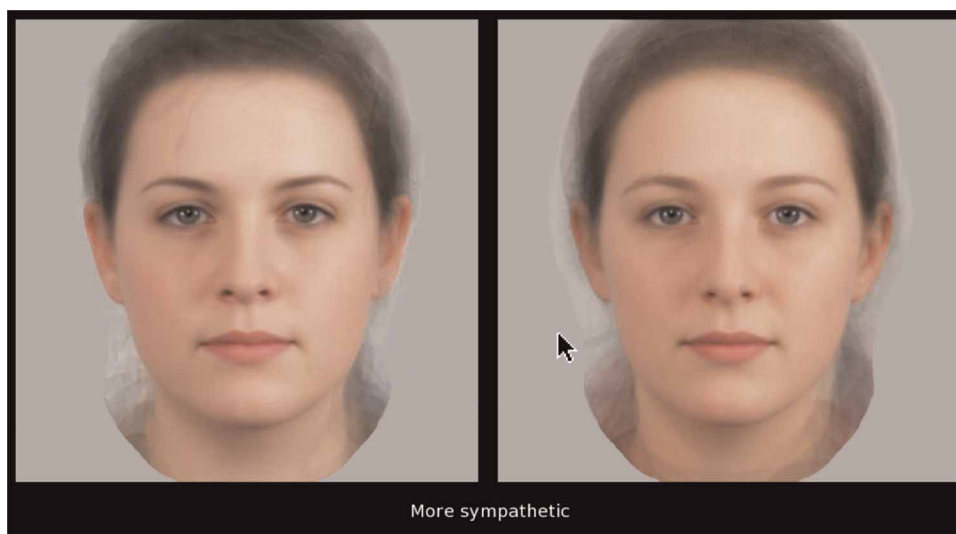


Figure 3. An example stimulus display. Participants clicked on the face that better matched the discrimination statement. To view a colour version of this figure, please see the online issue of the Journal.

contributions to subscale scores for the women in our stimulus pool. For the PCS, we used discrimination statements based on Items 1 (health is better), 2b (has greater difficulty climbing stairs), 3a (accomplishes less due to health problems), and 5 (pain interferes more with work). For the MCS, we used discrimination statements based on Items 4a (accomplishes less due to emotional problems), 4b (works less carefully due to emotional problems), 6a (feels more calm and peaceful), and 6c (more often feels downhearted and low). The order of face pairs and questions was randomized for each participant. The presentation of high and low composites was balanced for field of presentation, both for individual participants and for the four questions used to assess each trait.

Before beginning the rating exercise, each participant also completed a computerized version of the Mini-IPIP personality inventory and the SF-12.

Results and discussion

There were three main findings. First, we replicated previous results showing that many personality traits can be accurately judged from static facial features (Little & Perrett, 2007; Penton-Voak et al., 2006). Second, we found that physical

health is also reflected in static facial composites. Third, we found that the internal features can by themselves carry much of the information used for personality and health judgements, although there were some elaborations and exceptions to this general rule. We now consider these points in turn.

Figure 4 shows discrimination accuracy for each trait. For the most part, traits clustered into two sets: one clearly at chance levels (conscientiousness, mental health), and the other set well above chance (agreeableness, extraversion, neuroticism, and physical health, all p s < .001, in these cases with both full and internal faces). Intellect/imagination was an interesting exception: Identification was significantly below chance levels with full faces, $t(89) = 2.27$, $p = .025$, yet well above chance with internal features only, $t(40) = 4.93$, $p < .001$. Internal features alone therefore allowed for accurate discrimination for four of the Big Five personality traits (conscientiousness was the exception), as well as physical health.

Figure 5 focuses on the difference in accuracy for full and internal faces. This difference indicates the benefits or costs of external features on identification. As evident from the figure, there were three significant differences between full and internal composites. External features contributed to

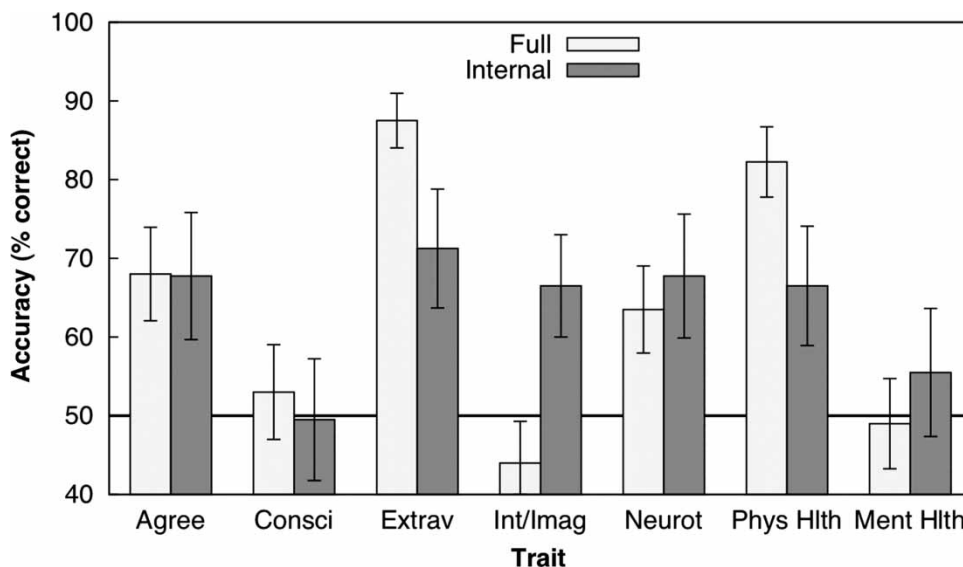


Figure 4. Accuracy on forced-choice (two-alternative) discrimination for the Big 5 personality traits and for physical and mental health, as measured by the appropriate subscales of the Short-Form 12-Item Health Survey (SF-12). Chance performance level is indicated by a line at 50%. Error bars indicate 95% confidence interval and can be used to compare conditions to baseline (i.e., error bars overlapping the 50% line are not significantly different from chance). Agree = agreeableness; Consci = conscientiousness; Extrav = extraversion; Int/Imag = intellect/imagination; Neurot = neuroticism; Phys Hlth = physical health; Ment Hlth = mental health.

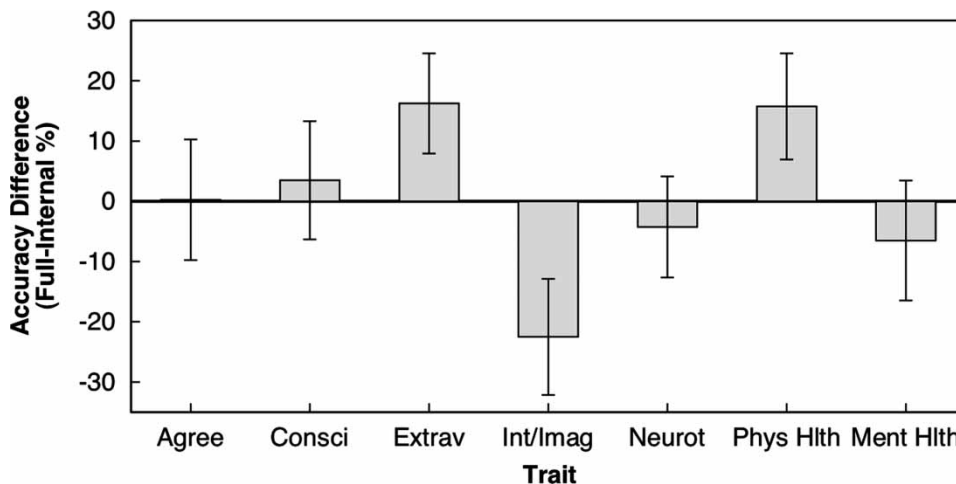


Figure 5. Difference in identification accuracy for full and internal faces. Positive bars indicate greater accuracy for full faces; negative bars indicate greater accuracy with internal features only. Error bars indicate 95% confidence intervals. Agree = agreeableness; Consci = conscientiousness; Extrav = extraversion; Int/Imag = intellect/imagination; Neurot = neuroticism; Phys Hlth = physical health; Ment Hlth = mental health.

accurate discrimination for both physical health and extraversion, $t_s(129) > 3.70$, $p_s < .001$. For physical health, the PCS subscale includes questions in which excess body weight might produce

lower scores (for example, “my health limits me in climbing several flights of stairs”). Inspection of Figure 1 shows that for full faces, there is evidence of additional body weight in the outline of the low

compared to high physical health faces. Cues to body weight from the face and jaw outline are not available in the internal faces, which emphasize the spatial relationships between facial parts.

Finally, there is the case of intellect/imagination, in which the external features actually produce systematic error in identification, compared to the internal features alone. Of the four questions on this subscale, two related to facility with abstract ideas and two to imagination. We ran a two-factor analysis of variance (ANOVA), looking at accuracy on intellect/imagination discrimination as a function of Face Type (full or internal) \times Question Type (abstract ideas or imagination related). External features produced interference on both estimates of imagination and abstract ideas, evidenced by the main effect of face type, $F(1, 129) = 24.32$, $p < .001$. However, the two-way interaction of Face Type \times Question Type was marginal, $F(1, 129) = 3.58$, $p = .061$, although the form of the interaction was such that there was little effect of question type for full faces, and the benefit for internal over full faces was greater with imagination than with intellect questions.

EXPERIMENT 2: PERCEIVED ATTRACTIVENESS AND HEALTH

Experiment 1 demonstrated accurate perception of physical health, extraversion, agreeableness, and neuroticism from the face. Internal features alone were sufficient for better than chance recognition of all these traits. However, external features, at least in combination with the internal ones, improved accuracy of physical health and extraversion judgements. In contrast, intellect/imagination, especially as tapped by imagination, was apparent from internal features, and external features were actually misleading.

Do these results reflect accurate discrimination of specific traits? Or is it possible that our results could be explained by a more general effect? The attractiveness “halo”, in which socially desirable traits are indiscriminately applied to attractive people, is one such effect (Dion, Berscheid, & Walster, 1972). The problem with any such

account is that an indiscriminate halo effect cannot by itself explain the main findings of Experiment 1—namely, the cases of accurate discrimination. That is, if socially desirable traits were assigned to faces in a genuinely indiscriminate way, identification accuracy would be at chance. However, to the extent that perceived attractiveness is correlated with actual trait measures, then responses based on attractiveness could produce correct identification. Suppose, purely for illustration, that attractive people simply had the socially desirable values of the traits that were accurately identified. That is, suppose that attractive women were more extraverted, more agreeable, less neurotic, and physically healthier than less attractive women. Observers could then perform well simply by assigning the more attractive face the more desirable trait. Other kinds of global characteristics might similarly collect socially desirable trait values. For example, given the potential importance of health in problems of mate choice, it might also have been the case that socially desirable traits covaried with healthy appearance. Given the theoretical interest in the relationship between attractiveness and health outlined earlier, and the importance of perceived health in other contexts (Kramer, Arend, & Ward, 2010), we were interested in a possible health “halo”, in which socially desirable traits might be attributed according to perceived health.

In this experiment, we looked at how discrimination performance in Experiment 1 related to the attractiveness of the different composites and to the perceptions of their physical health. However, it may be important to reemphasize that we are not investigating a type of halo effect in which raters are indiscriminately applying socially desirable traits to attractive people. Instead, we are looking at the possibility that people who are rated as attractive (or healthy looking) actually have socially desirable traits.

Method

Design

The experiment was defined by two factors describing the stimulus images: Trait (agreeableness,

conscientiousness, extraversion, neuroticism, intellect/imagination, physical health, mental health) \times Face Type (full face or internal features only). Face type was varied between participants, trait within.

Stimuli

The same images were used as those in Experiment 1.

Participants

As described previously, all participants from Experiment 1 took part in this experiment.

Procedure

Participants rated single face images for physical health and attractiveness, in separate blocks. Images were presented one at a time in the centre of the screen, the same size as in Experiment 1. Under the image would appear a reminder phrase indicating the task for that block (e.g., "How attractive is this face?"), and under that reminder, a written 7-item scale (e.g., very unattractive; unattractive; slightly unattractive; average; slightly attractive; attractive; very attractive). A similar scale was used for physical health ratings (very unhealthy; unhealthy; slightly unhealthy; average; slightly healthy; healthy; very healthy). We also included a similar block in which participants rated relationship preference for the face, but technical errors in presentation of the scale invalidated subsequent analysis. Participants clicked on the appropriate rating with the mouse, and the next image then appeared.

Blocks were presented in counterbalanced order across participants. Prior to each block of trials, an instruction screen appeared showing an array of the faces about to be rated and instructions on the rating task to be performed (e.g., "In this section you will be judging the ATTRACTIVENESS of the faces above. Please take a moment to consider the range of attractiveness in these faces.").

Results and discussion

We first wanted to confirm that ratings were equivalent for participants familiar and unfamiliar with

the women in the stimulus pool. We did not expect any difference, as in a composite of 15 faces, the identities of the individual faces seemed impossible to discern. Other reports have suggested that individual faces are effectively disguised within composites of even six faces (Little & Hancock, 2002). When we correlated the ratings given by the two groups to each of the 15 face stimuli, we found high correlations both for attractiveness, $r(13) = .91$, $p < .001$, and for physical health, $r(13) = .75$, $p = .001$. In addition, the two groups did not differ on attractiveness ratings, $t(14) = 0.27$, $p = .793$, or health ratings, $t(14) = 0.60$, $p = .556$. The ratings of the two groups were therefore combined in further analyses.

There was general agreement in the attractiveness ratings and, to a lesser degree, the health ratings given to the full and internal faces. For attractiveness, the correlation between the mean ratings given to each full and corresponding internal face image was $r(13) = .66$, $p = .008$; the agreement on health was marginal, $r(13) = .47$, $p = .077$.

However, our main focus is on the detailed relationship between accuracy of trait discrimination and the health and attractiveness ratings, as computed separately for full and for internal face images. Below, we present the complete results for each rating (attractiveness and health) and then selectively highlight interesting findings. To summarize, attractiveness and perceived health appear to be honest signals of extraversion and physical health, even when information is limited to internal face features. However, there are numerous cases in which discrimination performance and ratings of these signals appear to be independent.

Attractiveness

We first consider attractiveness ratings. Figure 6 shows the difference in the mean attractiveness rating given to the high and the low value composite for each trait. The difference is separately shown for full and internal face images.

To better understand the attractiveness differences illustrated in Figure 6, we compared differences in attractiveness with differences in

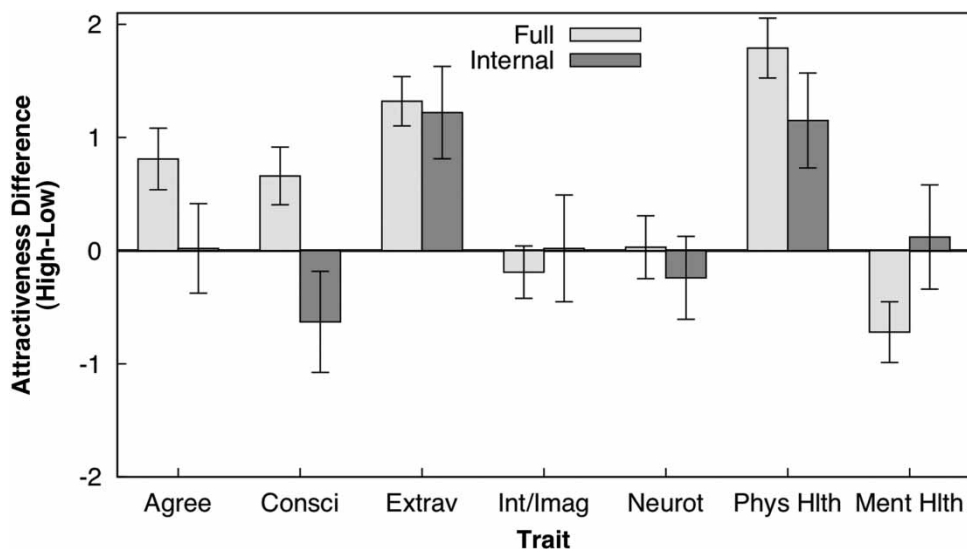


Figure 6. Difference in attractiveness for high and low trait composites. Attractiveness was rated on a 7-point scale from “very unattractive” (1) to “very attractive” (7) for all faces. Positive bars indicate greater attractiveness for high faces, negative bars indicate greater attractiveness for low faces. Error bars indicate 95% confidence intervals. Agree = agreeableness; Consci = conscientiousness; Extrav = extraversion; Int/Imag = intellect/imagination; Neurot = neuroticism; Phys Hlth = physical health; Ment Hlth = mental health.

discrimination accuracy in Experiment 1. These comparisons are shown in Table 1.

Table 1 separately summarizes the results for attractiveness and discrimination accuracy for the full and internal face stimuli. Beginning with the full faces (Table 1a), for three of the seven traits—agreeableness, extraversion, and physical health—

above-chance discrimination was accompanied by significant differences in attractiveness, such that attractive composites possessed the more socially desirable trait levels. However, this relationship between attractiveness and socially desirable traits did not hold across the board. Raters in Experiment 1 were unable to accurately discriminate

Table 1. Comparison of differences in attractiveness with differences in trait discrimination accuracy

Faces	Difference in attractiveness?	Discriminated trait levels?	
		Yes	No
a. Full	Yes	Extraversion, Physical health, Agreeableness	Conscientiousness, Mental health
	No	Neuroticism	Intellect/Imagination
b. Internal	Yes	Extraversion, Physical health	Conscientiousness
	No	Agreeableness, Neuroticism, Intellect/Imagination	Mental health

Note. Breakdown of traits according to whether high and low composites were accurately identified in Experiment 1 and whether there was a difference in attractiveness such that the more socially desirable trait pair was rated higher in attractiveness in Experiment 2.

conscientiousness, even though the same raters in the current experiment found the high conscientiousness face significantly more attractive than the low. Similarly, raters in Experiment 1 were unable to discriminate levels of mental health, even though they found the low mental health face more attractive than the high. The reverse pattern was also found. In Experiment 1, raters were able to discriminate levels of neuroticism in the full faces, even though in the present experiment both faces were rated equally attractive.

Attractiveness likewise does not provide a good explanation of discrimination for internal face images, summarized in Table 1b. As with the full faces, extraversion and physical health were accurately identified in Experiment 1 and also showed a significant difference in attractiveness. However, agreeableness, neuroticism, and intellect/imagination were also accurately identified from internal faces in Experiment 1, even though the high and low values of each were rated equally attractive. Finally, the low conscientiousness internal face was rated significantly more attractive than the high, but there was no accurate discrimination of these items in Experiment 1.

A consistent result therefore with both full and internal faces is that high levels of extraversion and physical health are reflected in attractive faces. In this context, it is also interesting to note that extraversion and physical health were the two traits that benefited significantly from information outside the internal faces (Figure 4). That is, there is information present in the full, and to a lesser extent, in just the internal facial features, which both is attractive to look at and serves as an honest signal of extraversion and physical health. However, attractiveness is not associated with all discriminable personality traits, and not all discriminable personality traits are reflected in corresponding attractiveness. The pattern of accurate performance in Experiment 1 is therefore not fully explained by an association of socially desirable traits and attractiveness.

While the above analysis investigates how attractiveness judgements at the group level relate to accurate trait perception, it does not address participants' decisions at the individual

level. We therefore carried out regression analyses to investigate whether differences in individual participants' ratings of attractiveness for the two composites (high minus low) predicted their accuracy—that is, did individual ratings predict subsequent discrimination? Of the seven traits for the full face judgements, only neuroticism accuracy was predicted by attractiveness ratings, $\beta = -.32$, $p = .014$ (Bonferroni corrected). For the internal faces, attractiveness did not predict accuracy for any of the traits. Again, this highlights attractiveness as unable to satisfactorily explain accuracy of perception in these judgements.

Health

A similar analysis was performed for perceived physical health (see Figure 7) and differences in perceived health compared with discrimination accuracy (Table 2). For the internal faces, health and attractiveness scores were also highly correlated, $r(13) = .70$, $p = .004$. However, with the sole exception of mental health, $t(40) = 1.63$, $p = .111$, all internal face pairs were perceived to reflect significantly different levels of physical health, all $t_s(40) > 2.60$, all $p_s < .013$. In these case, accurate discrimination of many traits could therefore conceivably be explained by a "health halo", such that healthy-looking people are not just perceived to have socially desirable traits, but actually do possess these traits. However, again, this cannot be a complete account of our findings—the trait pairs for conscientiousness differed in perceived health but could not be accurately discriminated.

Further exceptions to a perceived-health halo are found in the data for full faces. Health and attractiveness scores for the full faces were highly correlated, $r(13) = .91$, $p < .001$, and the general pattern of results is similar to that for attractiveness. Table 2a shows there were three cases in the full-face data in which socially desirable traits were both accurately discriminated and were also perceived as more healthy: agreeableness, extraversion, and actual physical health. But again, as with the attractiveness ratings, there were trait pairs that differ in perceived health but were not accurately discriminated (mental health) and trait pairs that were accurately discriminated but were

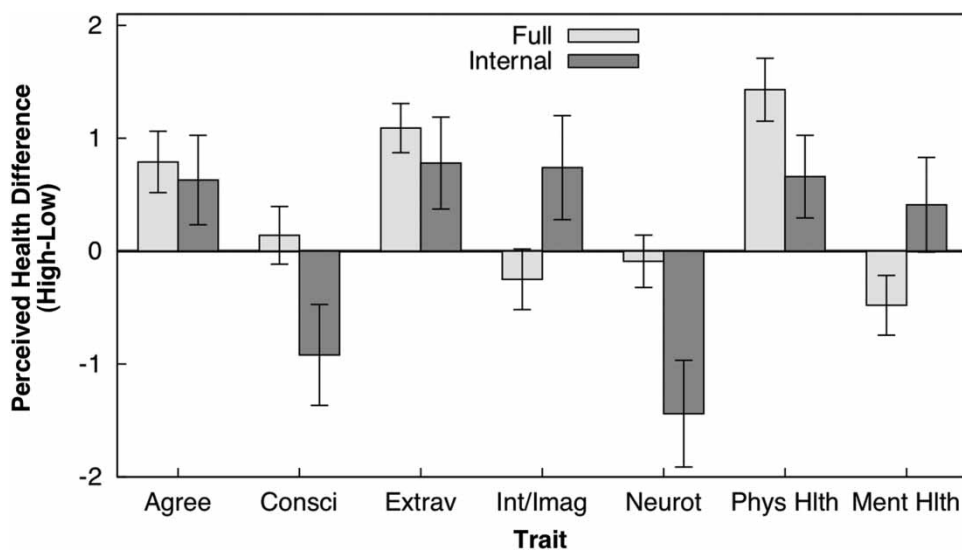


Figure 7. Difference in perceived health for high and low trait composites. Health was rated on a 7-point scale from “very unhealthy” (1) to “very healthy” (7) for all faces. Positive bars indicate greater perceived health for high faces; negative bars indicate greater perceived health for low faces. Error bars indicate 95% confidence intervals. Agree = agreeableness; Consci = conscientiousness; Extrav = extraversion; Int/Imag = intellect/imagination; Neurot = neuroticism; Phys Hlth = physical health; Ment Hlth = mental health.

Table 2. Comparison of differences in perceived health with differences in trait discrimination accuracy

Faces	Difference in perceived health?	Discriminated trait levels?	
		Yes	No
a. Full	Yes	Extraversion, Physical health, Agreeableness	Mental health
	No	Neuroticism	Conscientiousness, Intellect/Imagination
b. Internal	Yes	Extraversion, Physical health, Agreeableness, Neuroticism, Intellect/Imagination	Conscientiousness
	No	—	Mental health

Note: Breakdown of traits according to whether high and low composites were accurately identified in Experiment 1 and whether there was a difference in perceived health such that the more socially desirable trait pair was rated higher in perceived physical health in Experiment 2.

of equivalent apparent health (neuroticism). In summary, it seems unlikely that attractiveness or perceived health can explain all cases of accurate identification that we observed in Experiment 1.

As with attractiveness, we ran regression analyses with individual ratings of perceived health as a factor predicting trait accuracy. For full faces, perceived health differences predicted accuracy

for physical health only, $\beta = .38$, $p = .001$ (Bonferroni corrected). For internal faces, no differences in ratings predicted accuracy.

These results suggest that discriminations are not being made simply on the basis of halos relating to perceived attractiveness or apparent health. For example, neither attractiveness ratings at the group level, nor those at the level of individual raters, can well explain performance across all the different traits we have measured. Our results seem to dissociate raters' perceptions of attractiveness and healthy appearance from the ability of those raters to accurately judge personality traits. A related issue is to what extent having socially desirable traits is associated with attractiveness, independent of whether the trait can be accurately discriminated. For example, our raters could not discriminate levels of conscientiousness, even though the high conscientious composite was rated as more attractive than the low. Could it be the case that attractive, healthy-looking people will tend to have socially desirable traits, even if those traits are not accurately perceived by observers? Again, the relationship between attractiveness and socially desirable personality traits is not straightforward. For example, from our results it seems that attractive faces are more likely to reflect high than low levels of conscientiousness. However, neither attractiveness nor healthy appearance were associated with low neuroticism, or high intellect/imagination. These results suggest a more complex picture than any simple account relating a global measure such as attractiveness to social desirability for a multitude of personality traits.

GENERAL DISCUSSION

Previous work has shown that observers can accurately assess aspects of personality based on unfamiliar, static faces with neutral expressions (Little & Perrett, 2007; Penton-Voak et al., 2006). Our main results, from Experiment 1, show further that internal features of the face, specifically the areas around the eyes, nose, and mouth, carry enough information to allow accurate

judgements relating to physical health, and to four of the Big Five personality factors: agreeableness, extraversion, neuroticism, intellect/imagination (cf. openness). By comparing accuracy with full faces to internal features only, our method also allowed us to identify the contribution of external features (and colour) to identification. Although external features contributed to accurate identification of health and extraversion, they actually interfered with accurate judgements of intellect/imagination.

Experiment 2 verified that accuracy did not result from attractive people simply having more socially desirable traits. That is, the traits in which the composite pair differed in attractiveness were not necessarily correctly identified, and the traits that were correctly identified did not necessarily differ in the attractiveness of the composite pair. Likewise, our results do not seem completely consistent with the possibility that healthier looking people also have simply more socially desirable personality traits than less healthy looking people. Analyses of individual predictors further demonstrated that perceived health and attractiveness, while influencing judgements, did not account for accurate perceptions of personality.

As noted earlier, there were a few cases in which our composites overlapped in traits other than those they were created for. There was no surprise that MCS and neuroticism dimensions coincided for those individuals making up the composite pairs as these scales clearly reflect similar domains. That neuroticism but not MCS was discriminated from the images is more surprising, though this may simply reflect that the latter taps a more general domain that also includes depression, anxiety, and so on. Alternatively, MCS may just be a less well validated measure of mental health. In addition, the low agreeableness group had significantly lower conscientiousness than the high agreeableness group. This may mean that agreeableness composites were more easily discriminated because they differed on two trait dimensions. However, conscientiousness was not accurately discriminated, and so it seems unlikely that this extra information would have contributed significantly to participant accuracy.

Our results provide a useful replication of the Penton-Voak et al. (2006) and Little and Perrett (2007) findings. These studies used correlations of rated and actual traits, rather than the forced-choice identification we used. Little and Perrett used composite images, as we did, while Penton-Voak et al. also used individual images. These two studies asked observers to rate the degree of a trait (e.g., agreeableness present in the image), whereas we asked observers the same questions as those that were used to create the personality ratings. Despite these differences in images and observer tasks, all three studies found accurate identification of agreeableness and extraversion. Like Little and Perrett, we also find accurate identification of neuroticism, although in our internal as well as full-face images. Little and Perrett noted several potential advantages and disadvantages in the use of composites. One advantage is that traits consistently associated with specific visual features will have increased signal to noise ratio. The fact that neuroticism is, to date, found more easily with composite than single images suggests that the distinguishing visual characteristics for this trait are only weakly present in single images.

We have no compelling account as to why conscientiousness was accurately identified in the Little and Perrett (2007) images but not in ours. Similarly, there are differences in their study and ours in the attractiveness differences of composites. Little and Perrett only found differences between the high and low agreeableness female composites, whereas we found observed significant differences in extraversion and conscientiousness as well. This may simply reflect the reliability of trait differences across different samples but at present it is difficult to tell.

Although both Penton-Voak et al. (2006) and Little and Perrett (2007) investigated personality displays in male and female faces, our current research only explored female composites. This limitation was due to our sampling a population with a low number of males, and therefore making it impossible to produce sufficiently separate composites, and it is likely that the current results may differ to those expected from male

composites. Little and Perrett found that male composites only differed significantly for extraversion and suggest that males may contain fewer cues in the face to their actual personality than do females. This idea finds limited support in the literature, which has shown that women are believed to use more expressive and nonverbal behaviours than men (Briton & Hall, 1995) and are better nonverbal encoders of facial expressions than men (Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979). However, these relate to dynamic signals, and so further research is required to demonstrate their applicability to static features.

As noted above, the use of composites could potentially lose as well as gain signals. Some previous evidence suggests that fluctuating asymmetry (FA), in the face as well as the body, is a cue to developmental integrity and physical health (Thornhill & Møller, 1997). FA within a composite image will of course tend to be less than that in any of the components. However, although FA is an unlikely cue for physical health in the composites, physical health was still accurately identified in both full and internal face images. We also noted earlier that evidence of excess body weight is much reduced in the internal images. Skin blood colouration is also associated with perceived health (Stephen, Coetzee, Law Smith, & Perrett, 2009), with increased redness linked with higher levels of blood oxygenation, although this cue was also not available in the internal face images. While we accept that skin surface properties and FA probably play a role in assessing health, our results show that other features also signal health when these cues are minimal. At present, we suggest that health is signalled through a variety of cues, including FA and colour, but also the spatial arrangement of internal face features.

We close with some admittedly speculative, but perhaps intriguing, links between our results and theories of biological signal systems. In this context, we have seen that a signal, in this case, levels of socially desirable traits (such as high or low levels of agreeableness), are expressed on the face of the "sender" and are accurately detected by the "receiver" viewing the face. Theories of biological signal systems emphasize the perspectives of

both the sender and receiver: A signal must be sufficiently informative, sufficiently often, that receivers benefit from attending to it. That is, if a signal is uninformative or easily faked, there is no advantage or reason for the receiver to attend to it. Conversely, in a stable-state system, attention to a signal suggests that there is some net benefit to the receiver in attending. But this very validity opens the possibility of another selective pressure, for the sender to insert occasional deceptive messages, which benefit the sender, possibly at the expense of the receiver. That is, the receiver may be manipulated into acting against their own best interests (e.g., Dawkins & Krebs, 1978). What then keeps the system "honest"? For example, in the context of mate choice, an individual who could display false signals of exaggerated fitness might acquire a higher quality mate. In this context, the interests of the sender and receiver are not entirely opposing, but they are divergent, producing a pressure to exaggerate fitness. Why is it then that all faces do not express a socially desirable personality? A general conclusion from signal theory is that in such cases of divergent self-interests, a signal will generally not remain informative unless it entails costs that impact more heavily on less fit individuals ("costly signals"; Grafen, 1990; Zahavi, 1975). An interesting issue may therefore be identifying costs for expressing socially desirable traits on the face.

Original manuscript received 3 August 2009
Accepted revision received 23 February 2010
First published online 17 May 2010

REFERENCES

- Boothroyd, L. G., Jones, B. C., Burt, D. M., DeBruine, L. M., & Perrett, D. I. (2008). Facial correlates of sociosexuality. *Evolution and Human Behavior, 29*, 211–218.
- Briton, N. J., & Hall, J. A. (1995). Beliefs about female and male nonverbal communication. *Sex Roles, 32*, 79–90.
- Dawkins, R., & Krebs, J. R. (1978). Animal signals: Information or manipulation? In J. R. Krebs & N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (1st ed., pp. 282–309). Oxford, UK: Blackwell.
- Dion, K. L., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24*, 285–290.
- Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18*, 192–203.
- Fink, B., & Penton-Voak, I. S. (2002). Evolutionary psychology of facial attractiveness. *Current Directions in Psychological Science, 11*, 154–158.
- Gangestad, S. W., Thornhill, R., & Garver, C. E. (2002). Changes in women's sexual interests and their partners' mate retention tactics across the menstrual cycle: Evidence for shifting conflicts of interest. *Proceedings of the Royal Society of London B, 269*, 975–982.
- Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology, 144*, 517–546.
- Grammer, K., Fink, B., Møller, A. P., & Manning, J. T. (2005). Physical attractiveness and health: Comment on Weeden and Sabini (2005). *Psychological Bulletin, 131*, 658–661.
- Grammer, K., Fink, B., Møller, A. P., & Thornhill, R. (2003). Darwinian aesthetics: Sexual selection and the biology of beauty. *Biological Reviews, 78*, 385–407.
- Johnston, V. S., Hagel, R., Franklin, M., Fink, B., & Grammer, K. (2001). Male facial attractiveness: Evidence for a hormone-mediated adaptive design. *Evolution and Human Behavior, 22*, 251–267.
- Kalick, S. M., Zebrowitz, L. A., Langlois, J. H., & Johnson, R. M. (1998). Does human facial attractiveness honestly advertise health? Longitudinal data on an evolutionary question. *Psychological Science, 9*, 8–13.
- Kramer, R. S. S., Arend, I., & Ward, R. (2010). Perceived health from biological motion predicts voting behaviour. *The Quarterly Journal of Experimental Psychology, 63*, 625–632.
- Little, A. C., & Hancock, P. J. B. (2002). The role of masculinity and distinctiveness in judgments of human male facial attractiveness. *British Journal of Psychology, 93*, 451–464.
- Little, A. C., & Perrett, D. I. (2007). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology, 98*, 111–126.
- Mazur, A., & Booth, A. (1998). Testosterone and dominance in men. *Behavioral and Brain Sciences, 21*, 353–397.

- Penton-Voak, I. S., Perrett, D. I., Castles, D. L., Kobayashi, T., Burt, D. M., Murray, L. K., et al. (1999). Menstrual cycle alters face preference. *Nature*, *394*, 884–887.
- 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*, 490–524.
- Rhodes, G., Chan, J., Zebrowitz, L. A., & Simmons, L. W. (2003). Does sexual dimorphism in human faces signal health? *Proceedings of the Royal Society of London B (Suppl.)*, *270*, S93–S95.
- Rhodes, G., Zebrowitz, L. A., Clark, A., Kalick, S. M., Hightower, A., & McKay, R. (2001). Do facial averageness and symmetry signal health? *Evolution and Human Behavior*, *22*, 31–46.
- Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore: The Johns Hopkins University Press.
- Shevlin, M., Walker, S., Davies, M. N. O., Banyard, P., & Lewis, C. A. (2003). Can you judge a book by its cover? Evidence of self-stranger agreement on personality at zero acquaintance. *Personality and Individual Differences*, *35*, 1373–1383.
- Stephen, I. D., Coetzee, V., Law Smith, M., & Perrett, D. I. (2009). Skin blood perfusion and oxygenation colour affect perceived human health. *PLoS ONE*, *4*, 1–7.
- Swaddle, J. P., & Reiersen, G. W. (2002). Testosterone increases perceived dominance but not attractiveness of human males. *Proceedings of the Royal Society of London B*, *269*, 2285–2289.
- Thornhill, R., & Møller, A. P. (1997). Developmental stability, disease and medicine. *Biological Reviews*, *72*, 497–548.
- Ware, J. E., Jr., Kosinski, M., & Keller, S. D. (1996). A 12 item short form health survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care*, *34*, 220–233.
- Weeden, J., & Sabini, J. (2005). Physical attractiveness and health in Western societies: A review. *Psychological Bulletin*, *131*, 635–653.
- Zahavi, A. (1975). Mate selection: A selection for a handicap. *Journal of Theoretical Biology*, *53*, 205–214.