



Interpretation of Appearance: The Effect of Facial Features on First Impressions and Personality

Karin Wolffhechel^{1*}, Jens Fagertun², Ulrik Plesner Jacobsen¹, Wiktor Majewski¹, Astrid Sofie Hemmingsen³, Catrine Lohmann Larsen³, Sofie Katrine Lorentzen³, Hanne Jarmer¹

1 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark, **2** Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark, **3** Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark

Abstract

Appearance is known to influence social interactions, which in turn could potentially influence personality development. In this study we focus on discovering the relationship between self-reported personality traits, first impressions and facial characteristics. The results reveal that several personality traits can be read above chance from a face, and that facial features influence first impressions. Despite the former, our prediction model fails to reliably infer personality traits from either facial features or first impressions. First impressions, however, could be inferred more reliably from facial features. We have generated artificial, extreme faces visualising the characteristics having an effect on first impressions for several traits. Conclusively, we find a relationship between first impressions, some personality traits and facial features and consolidate that people on average assess a given face in a highly similar manner.

Citation: Wolffhechel K, Fagertun J, Jacobsen UP, Majewski W, Hemmingsen AS, et al. (2014) Interpretation of Appearance: The Effect of Facial Features on First Impressions and Personality. PLoS ONE 9(9): e107721. doi:10.1371/journal.pone.0107721

Editor: Nouchine Hadjikhani, Harvard Medical School, United States of America

Received: March 13, 2014; **Accepted:** August 20, 2014; **Published:** September 18, 2014

Copyright: © 2014 Wolffhechel et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and the Supporting Information files.

Funding: These authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* Email: karinw@cbs.dtu.dk

Introduction

We tend to evaluate others on their appearance and then move on to treat and interact with them based on these first impressions. Such an opinion can be formed after a tenth of a second from faces with neutral expressions and additionally people assess faces similarly for multiple traits, e.g. dominant and extraverted [1–3]. Specific facial features important for generating a first impression have been identified, for example is a large facial width-to-height ratio used as an indicator for a less trustworthy and more dominant personality type [4–6] - perhaps due to higher levels of testosterone in the blood resulting in a wider face [7]. Another important feature are the eyes; eye contact evokes trustworthiness [8]; and large eyes make a person appear more empathetic, agreeable, extraverted, conscientious and intelligent [9]. On one hand, there is some truth behind first impressions - it has been shown that valid inferences are made for at least four personality traits (Agreeableness, Conscientiousness, Extraversion, and Dominance) from facial features [10–13] - on the other hand, first impressions are not always accurate, e.g. people with infant-like facial traits (small chin, high eyebrows, and large eyes) are perceived as more emotionally warm, submissive, and naive [14], but often the direct opposite is true, as seen in many adolescent boys [15]. One cause of these inaccuracies is that people generate trait evaluations based on neutral facial expressions resembling actual emotional expressions - an effect named the overgeneralisation hypothesis [16].

To further delve into the generation and validity of first impressions, the differences and commonalities between faces have been studied extensively. An often-used approach is to go from a high-dimensional representation of a face, e.g. pixel values, 3-dimensional scans of faces or annotations of facial landmarks, to a lower-dimensional face-space by a Principal Component Analysis (e.g. [17]). Each dimension of the new face-space defines global properties of a face, which cannot be reduced to single features [3]. The implementation of such a face-space has made it possible to generate artificial faces supposedly expressing traits to a low or to a high degree [18,19]. Walker and Vetter [20] used this technique to manipulate photographs of real faces making them appear more extreme for a given trait. Validation of these changed faces showed them to be chosen slightly more often than their non-extreme counterparts. In 2011 Rojas *et al.* showed facial trait evaluations as predicted automatically from facial features with high accuracy, revealing the consensus between participants when rating a face [21].

Since trait evaluations are connected to a person's facial structure it was our focus to generate a more complete picture of the relationship between facial features, trait evaluations made by others and measured personality traits. Our results confirmed the importance of facial features for trait evaluations and additionally some interesting connections between self-measured personality traits and first impressions surfaced. The artificial faces visualising the extremes of all traits were generated for men and women separately.

Materials and Methods

Ethics Statement

Data collection and analysis was performed in accordance with the Act of Processing of Personal Data and approved by the Danish Data Protection Agency before the beginning of the project. The participants were asked to give verbal informed consent to participate in this study and no data was collected until this consent was given. The consent is thereby documented by the recording of the data. This was in accordance to the guidelines of the Danish National Ethics Committee which state that written consent is only required if biological samples are collected, which was not the case in this study.

Participants

Participants ($N = 244$, 128 women, 116 men) were recruited on campus at the Technical University of Denmark. All were either employed or studying at the university and between 18 years and 37 years old ($\mu = 24.56$, $\sigma = 3.24$).

Photographs

Facial photographs of all participants were taken with a Canon PowerShot XC200 camera under standardised conditions; controlled lighting, a white background and the same distance to the camera.

Questionnaire

Each participant was instructed to fill out an online, Danish questionnaire composed of twelve questions regarding specific traits for twenty other randomly chosen, unacquainted participants from the cohort. Nine of the twelve traits were chosen to cover the personality traits measured with a self-report questionnaire and additionally we added the traits attractiveness, masculinity and physical health due to their possible effect on the other trait evaluations. The questionnaire was set up as a 9-range Likert scale with a neutral answer corresponding to five. The participants were instructed to evaluate each face for the traits friendly, adventurous, temperamental, physically healthy, extravert, dominant, attractive, masculine, emotionally stable, responsible and intelligent. The questions were phrased as “How [trait] does this person look” with the response scale ranging from 1, “Not [trait] at all”, to 9, “Very [trait]”. There was no time constraint for answering the questions and the faces were presented in randomised order. Approximately twenty participants rated each participant and the mean of the scores for each question was used as the actual score for that participant. Calculation of the Cronbach’s α confirmed the reliability of this approach. The scores are further on referred to as the *Ratings*.

Personality measurements

Cubiks In-depth Personality Questionnaire, CIPQ 2.0, a normative self-report questionnaire scoring 17 personality traits covering the Big Five [22] was used to measure the participants’ personality traits. The personality traits were scored in a range from 1 to 10 and the results were assessed during a 45-minute session with a certified CIPQ test-scorer and the respondent. The test measures Neuroticism as its low pole: Emotional Stability. The scores are reversed compared to Neuroticism and additionally Emotional Stability focuses less on a person’s level of anxiety and stress, but more on how emotionally perceptive and sensitive a person is. 226 participants completed the questionnaire.

Appearance Model, AM

An Appearance Model, AM, which models all texture and shape information inside the boundaries of a face, was used to derive the facial components. Two models were generated, one for each gender, due to large differences in facial composition between men and women. The model is built by first annotating all photographs regarding the position and size of facial landmarks. Shape variations of the faces are extracted by a Principal Component Analysis, which in this case resulted in over 30 principal components each interpreting certain holistic facial characteristics of the participants (32 principal components for the male faces and 35 for the female faces). Afterwards the texture information is extracted through removing all shape information by warping all the images onto a mean shape. A Principal Component Analysis is performed on the pixel intensities in this set of shape-neutral images to model the variation in texture. This resulted in over 60 principal components explaining the texture of the faces (62 components for the male faces and 71 for the female faces). The resulting model contains a number of facial components describing a given face [23]. An example of two facial components and their interaction is shown in Figure 1.

Data processing

Calculations were performed in R [24] and figures generated with the packages ggplot2 [25], pheatmap [26] and Cytoscape version 2.8 [27]. Differences in scores were tested for statistical significance with a Welch’s t-test and the correlations between the individual *Ratings* and the self-measured personality traits were evaluated using the Pearson Correlation Coefficient, r . The significance of the correlations between the *Ratings* and the personality traits were confirmed by a permutation test, a statistical significance test, with 10,000 repeats. A permutation test repeatedly calculates correlations for randomised data to thereby find a measure for the significance of the actual correlation. A number of different models, both non-linear and linear, with varying subsets of facial features as predictors, were built for the prediction of the *Ratings* and the personality traits. The training

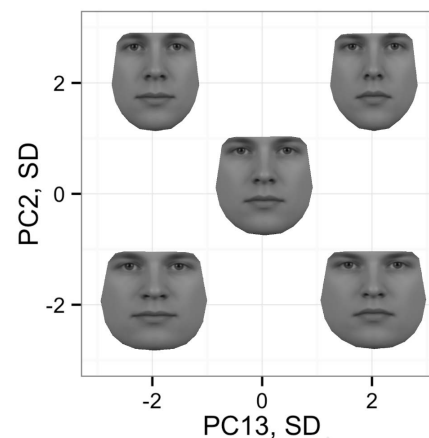


Figure 1. Example of two facial features, PC2 and PC13, and their interaction. The faces visualise how two principal components, PC, extracted by an Appearance Model, interact with each other. The coordinate system shows the change in a face when a principal component is moved two standard deviations in either the positive or the negative direction. The face in the middle shows the mean for all factors. E.g. the face in the upper right shows PC2 and PC13 at +2 standard deviations. It is seen that PC13 explains the shape of the mouth and PC2 the face width.

doi:10.1371/journal.pone.0107721.g001

was run as a 20-fold cross-validation and repeated thirty times for reliable standard deviations. Each model was trained on the training set using a growing number of the most correlated facial features (selected on the training set) as input. The Pearson Correlation Coefficient, r , between the observed and predicted Rating scores in the test set was used to evaluate the performance of each model. The best model was subsequently chosen based on its average performance on all folds.

Extreme faces

The β -coefficients for the relevant predictors from the linear regression model were used to generate two extreme faces for each gender and each Rating. High and low scoring faces were calculated by applying four standard deviations to either the positive or the negative direction of the facial features. This was done to simulate faces evaluated as belonging to the ends of the Rating scale.

Validation

The artificial, extreme faces were validated by asking 116 people, who were not part of the previous cohort, to choose from a set of four faces, which face they found to express a given personality trait the most. One of the four faces was the extreme face for the given Rating and the other three were randomly generated from the same parameter space. The validation was done as part of an open house at the Technical University of Denmark and therefore time constraints required us to only validate five traits for each gender. The traits were chosen to cover various aspects of the twelve *Ratings*. See Figure S4 for an example of a set of four images used in the validation.

Results

The Cronbach's α ranged from 0.63 to 0.92 for the *Ratings* for each gender, with only *Responsible* and *Emotionally Stable* for the female faces having values below 0.70. Agreement between participants was generally larger for the male faces ($0.80 < \alpha < 0.92$) than for the female faces ($0.63 < \alpha < 0.87$). We used the average score for each *Rating* as a more reliable measure, based on responses from multiple people, for how a face is assessed by others.

The *Ratings* were seen to fall into three clusters (Figure 2), which we named dominance-masculinity, attractiveness-health-extraversion, and trustworthiness-friendliness. The attractiveness cluster seemingly represents the halo effect (the hypothesis stating that attractive people are evaluated more positively regarding positively loaded personality traits [28,29]): High scores for Attractive clustered with high scores for Extraverted, Emotionally Stable, Physically Healthy, and Adventurous. We further discovered a clear link between scores for Dominating and Masculine for men ($r(114) = .73, p < .001$), which was in agreement with previous results [14,30]. We compared the *Ratings* between genders with a Welsh's t-test and found that women generally are perceived as more trustworthy ($p = 3.19 \times 10^{-5}$), responsible ($p = 4.40 \times 10^{-10}$) and attractive ($p = 6.35 \times 10^{-9}$), whereas men are seen as more emotionally stable ($p = 4.04 \times 10^{-6}$).

Connecting the participants' personality-trait scores to the *Ratings* for each gender revealed subtle, but significant correlations ($.20 \leq \text{abs}(r) \leq .32, p < .01$), which did not overlap between genders (see Figure 3). For men the most significant link was between evaluations for *Responsible* and the personality trait *Trusting*, a sub-trait of *Agreeableness* ($r(116) = .27$). Additionally we found a tendency that men with a more calm personality appear more friendly and extraverted ($r(116) = .20$). For women

the strongest link was between the evaluations for *Emotionally Stable* and the personality trait *Striving*, a sub-trait of *Conscientiousness* ($r(128) = .32$). Dominance was also for women linked to higher scores in the corresponding personality trait *Shaping* ($r(128) = .23, p < .01$). Higher scores for *Openness to Experience* followed higher evaluations for many *Ratings* including the traits *Adventurous* ($r(128) = .28$) and *Friendly* ($r(128) = .27$). We found no connection between participants self-reported personality traits and the scores they gave others in the *Ratings*.

Since effect sizes from correlated average scores can be inflated [31,32], we also correlated the raw scores given by each individual judge with the personality scores and then calculated averages and standard deviations for all these correlations based on individual judges. This resulted in effect sizes dropping below statistical significance (see Figure S1) with large standard deviations ($.31 < \sigma < .39$) revealing a substantial individual factor in trait evaluations. Since it was our goal to investigate subtle effects of facial features on trait evaluations and we wanted a more complete measure of the trait evaluations we continued with the averaged scores, but the above found group-based effect sizes should be noted as inflated on the individual level.

Next, we explored the possibility of predicting single personality traits either from a person's *Ratings* or from his or her facial features. However, diverse non-linear approaches and varying subsets of predictors could not predict the personality traits, revealing the correlations as not strong enough for a stable prediction. The performance, when comparing the predicted and the observed personality traits, was low ($r < .20, \text{RMSE} > 2.00$) and residual plots showed no satisfying fit.

The prediction of a person's *Ratings* from his or her facial features, however, gave more reliable results: it is to a certain extent possible to predict how a given person will be perceived based on his or her facial characteristics. We found a linear regression model to be most accurate, whereas more complex models (e.g. support vector machines with linear and radial kernels and a neural network with varying numbers of hidden nodes) did not improve the prediction significantly. The scores for Friendly for men were predicted with the highest accuracy ($r = .65, \sigma = 0.04$). Figure 4 visualises the correlation between observed and predicted scores for all *Ratings* for both genders with the corresponding Cronbach's α . We observed predictions being overall better for male faces ($p < .001$), which is in agreement with the higher values of Cronbach's α for these. The correlation between the Cronbach's α and the prediction accuracy was substantial ($r = .51, p < .02$), again confirming the importance of the agreement between raters for the validity of a given prediction [20].

To visualise the models we generated artificial faces predicted to express a given trait either to a high or a low degree. Three pairs of these extreme faces for each gender are shown in Figure 5 and all face pairs are shown in Figures S2 and S3. Our model is built from holistic features and therefore it is difficult to conclusively state much about specific parts of a face, but some differences stand out in the extreme pairs. For appearing friendly the mouth seems to have an impact: a wider mouth with neutrally or upwards pointed corners of the lips resulted in higher scores for friendliness (Figure 4B). The male extreme faces for Dominating (Figure 4C) reveal the effect of a wider face and a more pronounced eyebrow-ridge. For women the extreme faces for Adventurous (Figure 4D) indicate a positive impact of fuller lips and dark lashes (possibly eye make-up).

We performed a validation of our extreme faces by asking 116 people outside the original study to choose between four artificial faces the one that looked to possess a certain personality trait to the

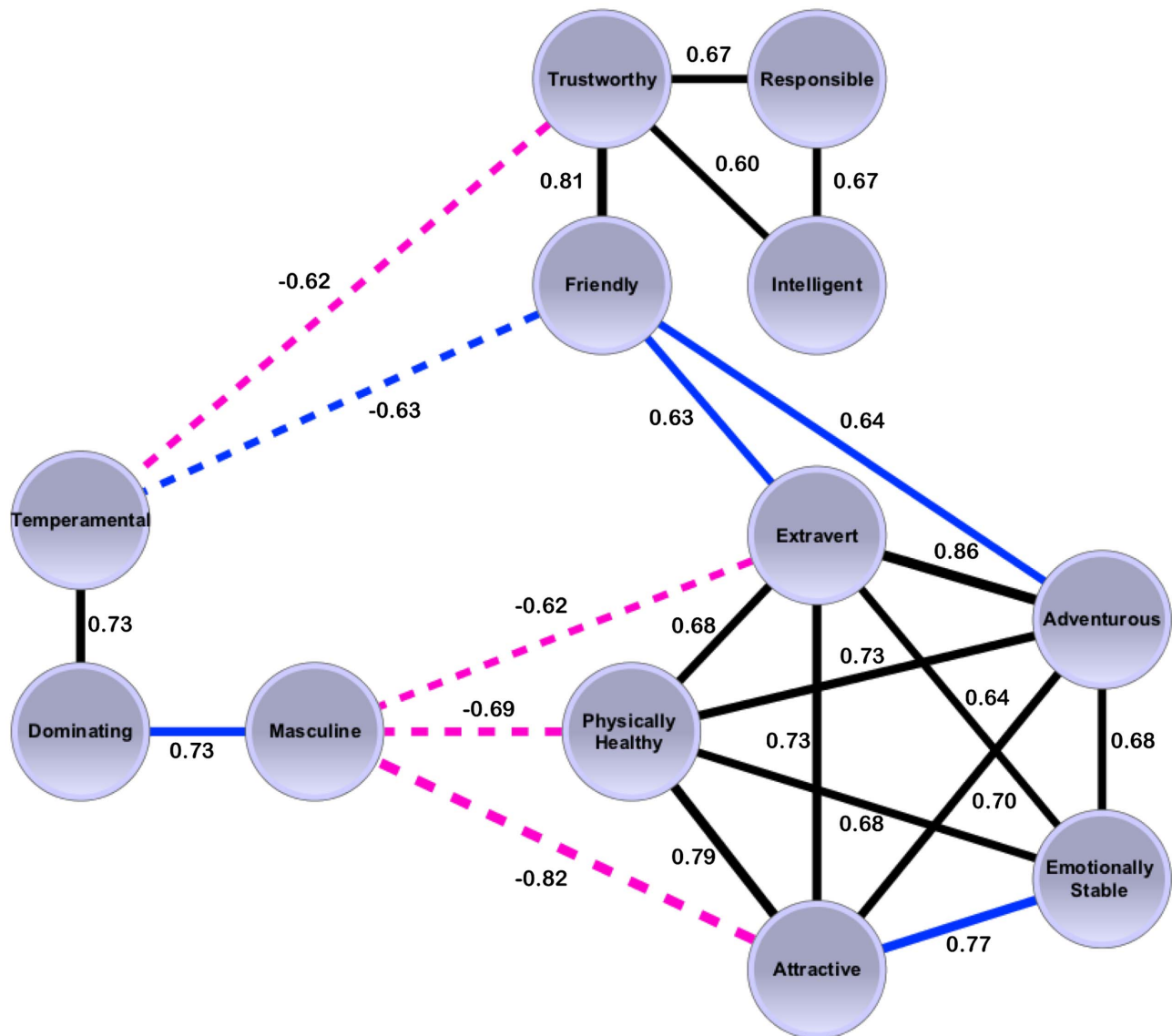


Figure 2. Network graph of all significant correlations between Ratings. The network depicts the relationship between the individual Ratings as the correlation coefficient, r , between scores. A dashed line depicts negative and a solid line positive correlations and the thickness of the line indicates the strength of the relationship with r as the edge label. Relationships significant for both genders are black, for men blue and for women magenta. Three clusters can be seen in the network with *Trustworthy*, *Responsible*, *Friendly* and *Intelligent* in the first, *Extraverted*, *Adventurous*, *Emotionally Stable*, *Attractive* and *Physically Healthy* in the second and *Temperamental*, *Dominating* and *Masculine* in the third. We named the clusters trustworthiness-friendliness, attractiveness-health-extraversion and dominance-masculinity. doi:10.1371/journal.pone.0107721.g002

highest degree; one of the four faces was the extreme face for the given trait. Due to time constraints this was done for only five traits for each gender and we saw that the extreme face was chosen above random in all cases except one, *Intelligent* for women. This unsuccessful case could be connected to a lower agreement between raters when evaluating this trait ($\alpha = 0.70$). In Figure 6 we show the percentage of times each face was chosen for each question and how this compares to a random selection between the four faces. Validation of the male extreme faces was successful: participants selected the extreme male face significantly more often than the random faces for the five traits ($p < .001$), which fits the fact that the prediction of the Ratings for the male faces had an overall good performance and the higher agreement between raters for the male faces. Validation of the female extreme faces

was significant for only two of the five chosen traits, *Friendly* and *Adventurous* ($p < .03$). The other three extremes for the traits *Dominating*, *Responsible* and *Intelligent* were not chosen significantly more often than the random faces, which is in concordance with the lower prediction performance, especially for *Responsible* and *Intelligent*.

Discussion

We found the prediction of personality traits from facial features to be unsuccessful, but we discovered that some traits could be inferred from a face to a certain extent. The identified connections between individual personality traits and Ratings were subtle but significant and mostly in accordance with previous research. For women we confirmed that inferences could be made about the

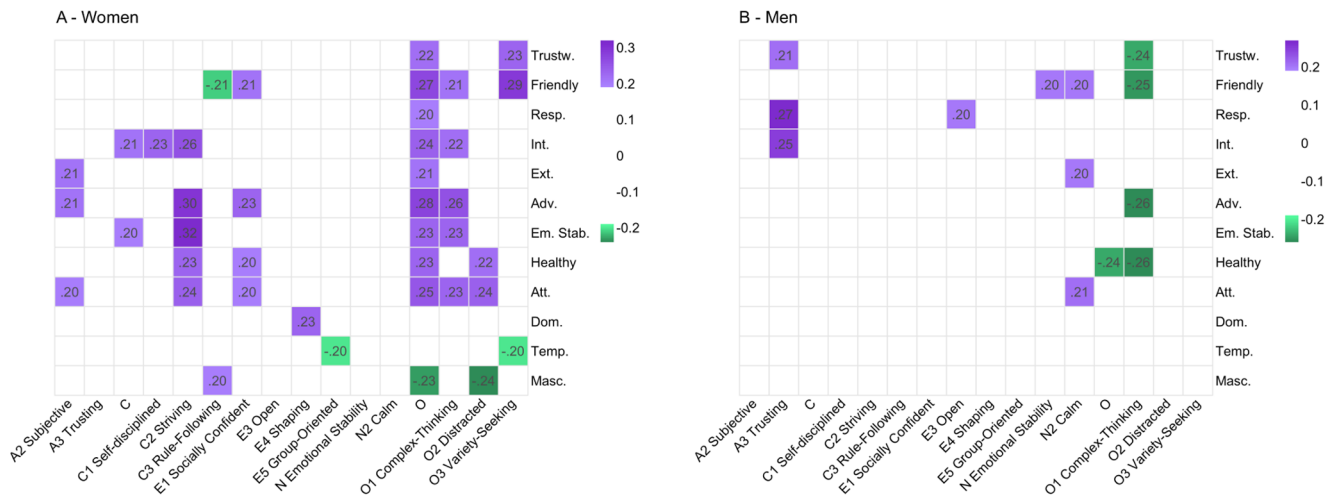


Figure 3. Correlations between Ratings and self-reported personality traits visualised by heat maps. Heat map A shows the correlations for women and heat map B the correlations for men. The personality traits are on the x-axis and the Ratings on the y-axis and a positive correlation is indicated with purple and a negative with green, where darker colours stand for bigger effect sizes. Only significant correlations with $abs(r) \geq .20$ and $p < .01$ are shown. Calculating the average of the correlations between personality traits and Ratings given by individual judges resulted in a drop in effect size; therefore the correlations in these heat maps should not be seen as significant on the individual level. Abbreviations for the Ratings are: Trustw. = Trustworthy, Adv. = Adventurous, Temp. = Temperamental, Healthy = Physically Healthy, Ext. = Extraverted, Dom. = Dominating, Att. = Attractive, Masc. = Masculine, Em. Stab. = Emotionally Stable, Resp. = Responsible and Int. = Intelligent. doi:10.1371/journal.pone.0107721.g003

level of *Openness*, *Striving* and *Dominance* from a face. For men a calmer personality linked with higher evaluations for friendliness and a more trusting personality with higher scores for responsibility. The latter could indicate that a person’s level of trust in others can be influenced by his or her appearance, if appearance makes others treat him or her as more responsible. Some of these correlations have been reported previously [10,11,13], but seldom from only facial photographs as input [12,34].

The above results were based on average Ratings; correlations for Ratings given by individual judges were significantly smaller and in some cases shrunk down to zero. Thus, when assessing individual scores for a connection between trait evaluations and

personality, an effect does not seem apparent: none of the average correlations on the individual level were statistically significant. We have two arguments for why this happens. Firstly, the individual correlations are only based on responses for about ten faces, since each judge rated approximately ten people of each gender, which leads to a much higher uncertainty in the correlation. Secondly, if there is a real connection between a trait and facial appearance, then based on classical test theory the averaging of several scores can reduce the trait evaluation error. This happens because each score is composed of a true component and an error component, leading to a decrease in error when scores from several raters are combined. Consequently we still see the average Rating score as

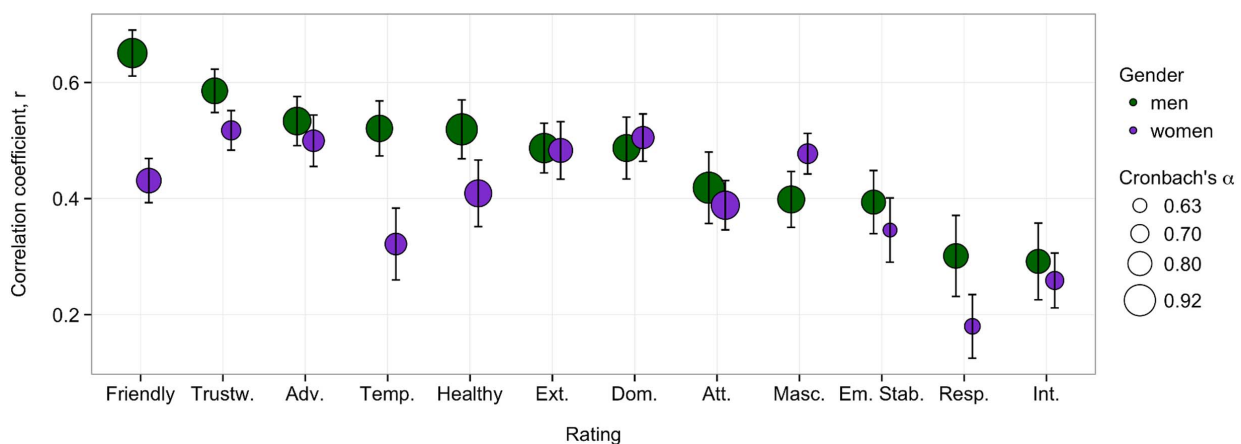


Figure 4. Prediction of Ratings from facial features. The plot shows the average correlation coefficient and standard deviation between observed and predicted scores for each Rating and each gender. A linear regression model was built in a 20-fold cross-validation with a varying number of the most correlated facial components as predictors, chosen based on the training set. Standard deviations are gathered by running the calculations thirty times with different folds for each run. The Ratings are in the plot ordered based on performance for the male faces. The size of the points indicates the Cronbach’s α for that trait and it is seen that larger α -values correlate positively with prediction performance. Abbreviations for the Ratings are: Trustw. = Trustworthy, Adv. = Adventurous, Temp. = Temperamental, Healthy = Physically Healthy, Ext. = Extraverted, Dom. = Dominating, Att. = Attractive, Masc. = Masculine, Em. Stab. = Emotionally Stable, Resp. = Responsible and Int. = Intelligent. doi:10.1371/journal.pone.0107721.g004

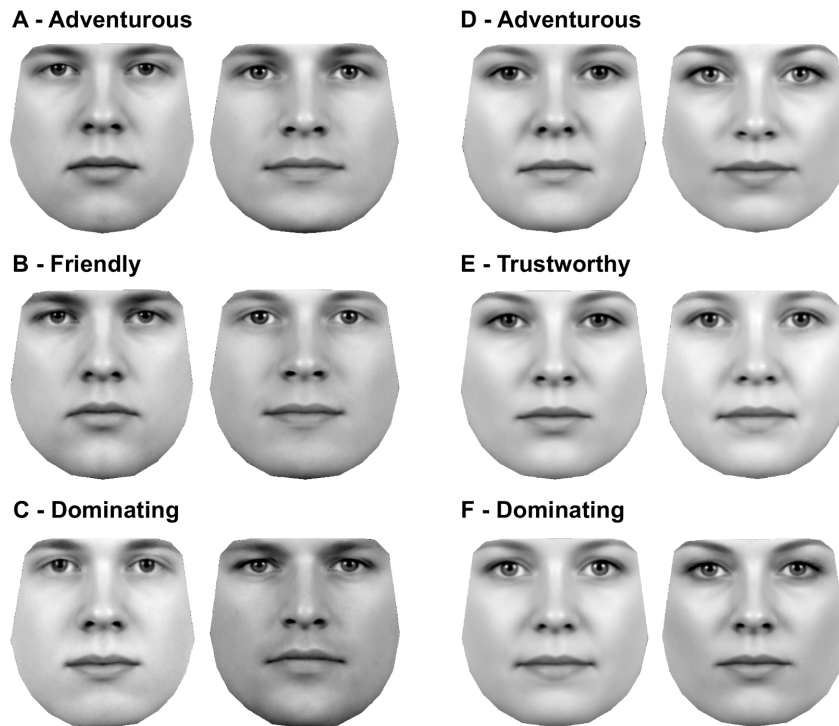


Figure 5. Extreme faces for the Ratings. For each face pair the left extreme face is predicted as being judged very low for a given trait and the right face as very high. Each face is based on the β -coefficients from the best linear regression model for that given Rating and gender. We generated the faces by multiplying each β -coefficient to either +4 standard deviations or -4 standard deviations of the matching facial component. A: Male extremes for *Adventurous*. B: Male extremes for *Friendly*. C: Male extremes for *Dominating*. D: Female extremes for *Adventurous*. E) Female extremes for *Trustworthy*. F: Female extremes for *Dominating*. doi:10.1371/journal.pone.0107721.g005

reliable for assessing trait evaluations from facial features, although the found correlations between personality traits and trait evaluations should be seen as only valid on a group-based level.

As others, we confirmed that people evaluate faces similarly for several traits, which manifested itself in a fairly accurate prediction of how people perceive a face based on facial features (e.g. [1]). This effect was supported by a validation of our generated extreme faces. The validation success was seen to be somewhat dependent on agreement between raters: traits with higher Cronbach's α were generally predicted with higher accuracy. For some traits, e.g. *Responsible* and *Intelligent* for women, the reliability of judgments was low, which revealed these traits as subjectively evaluated. In general raters agreed more on how to evaluate male faces.

The extreme faces confirmed the impact of a larger facial width-to-height ratio for appearing more dominating [4]. The shape of the mouth was also seen to have an impact, with neutral or upwards pointed corners of the lips resulting in higher scores for positively loaded traits. This could specifically be due to the overgeneralisation hypothesis leading to false trait judgments, since a more smiling expression connects well with emotional expressions for positive traits [16,20].

The *Ratings* were seen to fall into three clusters, dominance-masculinity, trustworthiness-friendliness, and attractiveness-health-extraversion. These three clusters fit previous findings showing three factors as sufficient for evaluating a face. Two of these factors, valence/trustworthiness and dominance, were discovered

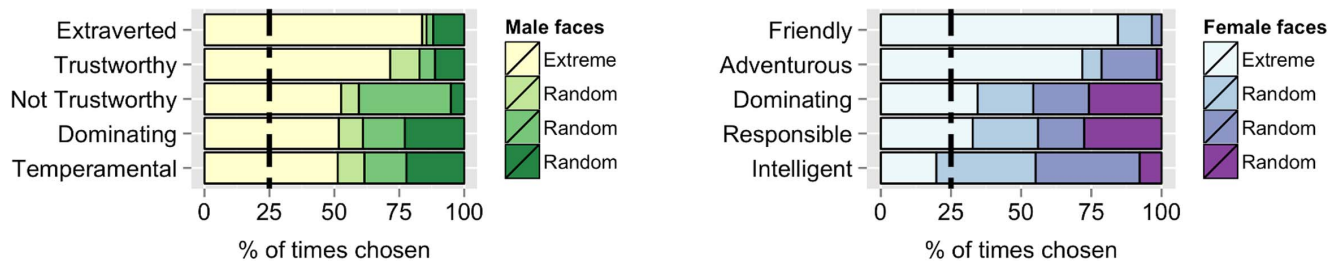


Figure 6. Validation of extreme faces. In the validation we presented four faces to 116 persons and asked them to choose which one they found to represent a given trait the most. The left plot shows results for the male extremes and the right results for the female extremes. The length of each section in each bar indicates the percentage of times the given face was chosen. The dotted line indicates the percentage representing a random selection of the extreme face. In all cases except one the extreme face was chosen more often than random. For the male faces we found the extremes to be chosen significantly over random. For the female faces this was only found for the *Friendly* and *Adventurous* extremes. The colours are from [33]. doi:10.1371/journal.pone.0107721.g006

by Oosterhof and Todorov [18] and confirmed by Sutherland *et al.* [35]. Sutherland *et al.* additionally detected a youthful-attractiveness factor, which connects to our attractiveness-health-extraversion cluster. These studies also found emotional expressions like anger and happiness to correlate strongly with trust evaluations, an effect also apparent in our extreme faces.

Perhaps the fact that faces are assessed based on the over-generalisation hypothesis led to us not finding a clear relationship between trait evaluations and self-measured personality scores. Studies using short video sequences of a person instead of a photograph have reported more precise first impressions [36], strengthening the belief of a connection between personality and appearance. It seems that a single facial photograph lacks information for evaluating diverse traits: a viewer will miss additional cues for gathering a more complete first impression from a face and will therefore instead focus overly on facial expressions.

In conclusion our results confirm the impact facial features have on first impressions and that people generally agree on how to evaluate some aspects of personality based on a face, even though these evaluations often are far from the self-measured personality traits. We replicated previous findings about three factors being sufficient for trait evaluations. We believe that appearance has an impact on personality development, since social interactions are such a monumental part of our lives. A slight indication of this was also found in some of the connections between self-reported personality traits and trait evaluations and in studies involving video sequences of ratees, but more research is at the moment needed to prove the directionality and size of this effect.

Supporting Information

Figure S1 Heat maps for the averaged correlations between Ratings given by individual judges and the self-reported personality traits. Heat map A shows the correlations for women and heat map B the correlations for men as a 95% confidence interval. The personality traits are on the x-axis and the Ratings on the y-axis and a positive correlation is indicated with purple and a negative with green, where darker colours stand for bigger effect sizes. Only the correlations significant in the correlated averages in Figure 3 are shown and a large drop in effect size is seen compared to these. Abbreviations for the Ratings are: Trustw. = Trustworthy, Adv. = Adventurous, Temp. = Temperamental, Healthy = Physically Healthy, Ext. = Extraverted, Dom. = Dominating, Att. = Attractive,

References

- Willis J, Todorov A (2006) First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science* 17: 592–598.
- Bar M, Neta M, Linz H (2006) Very First Impressions. *Emotion* 6: 269–278.
- Todorov A, Oosterhof NN (2011) Modeling Social Perception of Faces. *IEEE Signal Processing Magazine* 28: 117–122.
- Carré JM, McCormick CM, Mondloch CJ (2009) Facial Structure Is a Reliable Cue of Aggressive Behavior. *Psychological Science* 20: 1194–1198.
- Haselhuhn MP, Wong EM (2012) Bad to the bone: facial structure predicts unethical behaviour. *Proceedings Of The Royal Society* 279: 571–576.
- Haselhuhn MP, Wong EM, Ormiston ME (2013) Self-Fulfilling Prophecies as a Link between Men's Facial Width-to-Height Ratio and Behavior. *PLoS ONE* 8: e72259.
- Lefevre CE, Lewis GJ, Perrett DI, Penke L (2013) Telling facial metrics: facial width is associated with testosterone levels in men. *Evolution and Human Behavior* 34: 273–279.
- Bayliss AP, Tipper SP (2006) Predictive Gaze Cues and Personality Judgments: Should Eye Trust You? *Psychological Science* 17: 514–520.
- Paunonen SV, Ewan K, Eathly J, Lefave S, Goldberg H (1999) Facial Features as Personality Cues. *Journal of Personality* 67: 555–583.

Masc. = Masculine, Em. Stab. = Emotionally Stable, Resp. = Responsible and Int. = Intelligent.
(TIFF)

Figure S2 Male extremes for the Ratings. The extreme face scoring low for a given trait is depicted on the left and the extreme face scoring high on the right for each Rating. The traits are ordered based on prediction performance.
(TIFF)

Figure S3 Female extremes for the Ratings. The extreme face scoring low for a given trait is depicted on the left and the extreme face scoring high on the right for each Rating. The traits are ordered based on the prediction performance for the male faces.
(TIFF)

Figure S4 Example of a validation question, Intelligent. The upper left face is the generated extreme for the trait *Intelligent*. The other three are randomly generated from the same parameter space as the extreme face. The extreme face for *Intelligent* was the only one that was not chosen over random in the validation, which matched the fact that *Intelligent* also was predicted with the lowest performance.
(TIFF)

File S1 Supporting Information. Table S1, questions used for the Ratings. Table S2, participant information. Table S3, scores given in the Ratings. Table S4, PCA scores for women with participants in columns. Table S5, PCA scores for men with participants in columns. Table S6, validation questions. Table S7, validation participant information. Table S8, scores given in the validation.
(ZIP)

Acknowledgments

We would like to thank all participants for taking part in this study. We thank Peter Meinicke for help in setting up the photo studio and we thank Peter Hartmann for answering questions about the personality measurements. We thank Cubiks, in particular Sune Malmgren and Anne Broe Andersen for letting us use the CIPQ 2.0 Personality Test and for giving advice about the results.

Author Contributions

Conceived and designed the experiments: KW UPJ WM SKL HJ. Performed the experiments: KW JF UPJ WM ASH CLL SKL HJ. Analyzed the data: KW HJ. Contributed reagents/materials/analysis tools: JF UPJ. Wrote the paper: KW HJ.

- Beer A, Watson D (2008) Personality Judgment at Zero Acquaintance: Agreement, Assumed Similarity, and Implicit Simplicity. *Journal of Personality Assessment* 90: 250–260.
- Carney DR, Colvin CR, Hall JA (2007) A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality* 41: 1054–1072.
- Kramer RS, King JE, Ward R (2011) Identifying personality from the static, nonexpressive face in humans and chimpanzees: evidence of a shared system for signalling personality. *Evolution and Human Behavior* 32: 179–185.
- Little AC, Perrett DI (2007) Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology* 98: 111–126.
- Pivonkova V, Rubesova A, Lindova J, Havlicek J (2011) Sexual Dimorphism and Personality Attributions of Male Faces. *Archives of Sexual Behavior* 40: 1137–1143.
- Zebrowitz LA, Montepare JM (2008) Social Psychological Face Perception: Why Appearance Matters. *Soc Personal Psychol Compass* 2: 1497–1517.
- Zebrowitz LA, Kikuchi M, Fellous JM (2010) Facial Resemblance to Emotions: Group Differences, Impression Effects, and Race Stereotypes. *Journal of Personality and Social Psychology* 98: 175–189.

17. Brahnam S, Chuang CF, Shih FY, Slack MR (2005) Machine recognition and representation of neonatal facial displays of acute pain. *Artificial Intelligence in Medicine* 36: 211–222.
18. Oosterhof NN, Todorov A (2008) The functional basis of face evaluation. *PNAS* 105: 11087–11092.
19. Kleisner K, Chvátalová V, Flegr J (2014) Perceived Intelligence is Associated with Measured Intelligence in Men but Not Women. *PLoS ONE* 9: e81237.
20. Walker M, Vetter T (2009) Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision* 9: 1–13.
21. Rojas M, Masip D, Todorov A, Vitria J (2011) Automatic Prediction of Facial Trait Judgments: Appearance vs. Structural Models. *PLoS ONE* 6: e23323.
22. Costa PT, McCrae RR (1995) Domains and facets - hierarchical personality assessment using the revised NEO personality inventory. *Journal of Personality Assessment* 64: 21–50.
23. Cootes TF, Taylor CJ (2001) Statistical models of appearance for medical image analysis and computer vision. *Medical Imaging* 4322: 236–248.
24. R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available: <http://www.R-project.org/>.
25. Wickham H (2009) *ggplot2: elegant graphics for data analysis*. Springer New York. Available: <http://had.co.nz/ggplot2/book>. Accessed 2014 Aug 29.
26. Kolde R (2013) pheatmap: Pretty Heatmaps. Available: <http://CRAN.R-project.org/package=pheatmap>. R package version 0.7.7.
27. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27: 431–432.
28. Dion K, Berscheid E, Walster E (1972) What Is Beautiful Is Good. *Journal of Personality and Social Psychology* 24: 285–290.
29. Little AC, Burt DM, Perrett DI (2006) What is good is beautiful: Face preference reflects desired personality. *Personality and Individual Differences* 41: 1107–1118.
30. Prentice DA, Carranza E (2002) What Women and Men Should Be, Shouldn't Be, Are Allowed to Be, and Don't Have to Be: The Contents of Prescriptive Gender Stereotypes. *Psychology of Woman Quarterly* 26: 269–281.
31. Monin B, Oppenheimer DM (2005) Correlated Averages Vs. Averaged Correlations: Demonstrating The Warm Glow Heuristic Beyond Aggregation. *Social Cognition* 23: 257–278.
32. Brand A, Bradley MT (2012) More Voodoo Correlations: When Average-Based Measures Inflate Correlations. *The Journal of General Psychology* 139: 260–272.
33. Brewer CA. Colors from www.ColorBrewer.org. Available: <http://www.ColorBrewer2.org>. Accessed 2014 Jun 3.
34. Penton-Voak IS, Pound N, Little AC, Perrett DI (2006) Personality Judgments From Natural And Composite Facial Images: More Evidence For A “Kernel Of Truth” In Social Perception. *Social Cognition* 24: 607–640.
35. Sutherland CAM, Oldmeadow JA, Santos IM, Towler J, Burt DM, *et al.* (2013) Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition* 127: 105–118.
36. Borkenau P, Liebler A (1992) Trait Inferences: Sources of Validity at Zero Acquaintance. *Journal of Personality and Social Psychology* 62: 645–657.