Tilburg University

Which facial features are central in impression formation?

Jaeger, Bastian; Jones, Alex L.

Publication date:
2020

Document Version
Early version, also known as pre-print

Which Facial Features Are Central in Impression Formation?

Bastian Jaeger

Tilburg University

Alex L. Jones

Swansea University

Draft version: 6 September 2020

This paper is currently undergoing peer review. Comments are welcome.

Author Note

Bastian Jaeger, Department of Social Psychology, Tilburg University, The Netherlands; Alex L. Jones, Department of Psychology, Swansea University, UK.

We thank Iris Holzleitner, Anthony Lee, Amanda Hahn, Michal Kandrik, Jeanne Bovet, Julien Renoult, David Simmons, Oliver Garrod, Lisa DeBruine, and Benedict Jones for sharing the R code for their article “Comparing theory-driven and data-driven attractiveness models using images of real women’s faces”, which was used for some analyses reported in this article.

Correspondence concerning this article should be addressed to Bastian Jaeger, Department of Social Psychology, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: bxjaeger@gmail.com.

Abstract

Which facial characteristics do people rely on when forming personality impressions from faces? Previous research has uncovered an array of facial features that influence people's impressions. Even though some (classes of) features, such as facial width-to-height ratio or resemblances to emotional expressions, play a central role in theories of social perception, their relative importance in impression formation remains unclear. Here, we model faces along a wide range of theoretically important dimensions. We use machine learning to test how well 31 features predict impressions of trustworthiness and dominance in a diverse set of 597 faces. In line with overgeneralization theory, emotion resemblances were most predictive of both traits. Other features that have received a lot of attention in the literature, such as facial width-to-height ratio, were relatively uninformative. Our results highlight the importance of modeling faces along a wide range of dimensions to elucidate their relative importance in impression formation.

Keywords: social perception; personality impressions; emotional expressions; facial width-to-height ratio; machine learning

Which Facial Features Are Central in Impression Formation?

People spontaneously judge others' personality based on their facial appearance (Todorov et al., 2015). For example, impressions of trustworthiness and dominance—which represent fundamental dimensions on which faces are evaluated (B. C. Jones et al., 2019; Oosterhof & Todorov, 2008)—are formed within a few hundred milliseconds (Willis & Todorov, 2006). These impressions can be extremely consequential as they guide many important decisions, such as voting, criminal sentencing, and personnel selection (Olivola et al., 2014; Todorov et al., 2015). Which facial characteristics do people rely on when forming personality impressions from faces? Previous research has produced a long list of facial features that influence personality impressions (Hehman et al., 2019; Todorov et al., 2015). These findings on the basis of personality impressions also provide the foundation for broader theories in social perception aiming to explain the accuracy and functional significance of personality impressions (Carré et al., 2009; Todorov et al., 2008; Zebrowitz, 2017).

One class of characteristics that has received a lot of attention is the structural resemblance between a person's facial features and emotional expressions. Resting faces that resemble an expression of happiness (e.g., slightly upturned corners of the mouth) are perceived as trustworthy, whereas resting faces that resemble an expression of anger (e.g., lowered eyebrows) are perceived as dominant (Adams et al., 2012; Said et al., 2009). These findings are highlighted by overgeneralization theory, which aims to explain the functional significance of personality impressions and the cognitive mechanisms underlying impression formation (Todorov et al., 2008; Zebrowitz, 2017). Overgeneralization theory posits that, due to their relevance for interpersonal behavior, people are particularly attuned to detecting emotional expressions from faces. This sensitivity causes people to perceive emotion expressions (and associated traits) in faces that *merely resemble* an emotional expression. Thus, overgeneralization theory posits that perceived resemblances to emotional expressions are an important input in impression formation and, more generally, that personality impressions are caused by an oversensitive emotion detection system.

Other theories have focused on different features in impressions formation. For example, facial width-to-height ratio (fWHR) influences impressions of trustworthiness and dominance (Ormiston et al., 2017; Stirrat & Perrett, 2010, 2012; Valentine et al., 2014). Moreover, some have argued that fWHR is an indicator of various behavioral tendencies, such as aggression, because biological factors (e.g., testosterone) influence both facial morphology and behavioral dispositions (Carré, McCormick, & Mondloch, 2009; for counterarguments, see Kosinski, 2017; Wang, Nair, Kouchaki, Zajac, & Zhao, 2019). Thus, this perspective posits that fWHR is an important input in impressions formation and, more generally, that personality impressions can be accurate because facial appearance and behavioral dispositions have a common underlying cause.

The Importance of Different Facial Characteristics

Even though some facial characteristics (e.g., resemblances to emotional expressions, fWHR) occupy a more central role in theories in social perception, evidence on their actual *relative* importance for impressions formation remains sparse. To examine the importance of different features, previous studies have predominantly examined how one or a few features affect personality judgments. However, this approach has two important limitations.

First, many facial characteristics are correlated, making it difficult to isolate their unique effects (A. L. Jones, 2019). For example, resemblances to emotional expressions are correlated with a variety of other features such as facial width-to-height ratio (Deska et al., 2018), babyfacedness (Sacco & Hugenberg, 2009), and race (Bijlstra et al., 2014). Even when one dimension of interest is manipulated (e.g., resemblance to a happy facial expression), perceptions of other dimensions (e.g., babyfacedness) will also change. This raises the question whether personality impressions are indeed best explained by emotion resemblances, or rather by other classes of features that are related to emotion resemblances.

Second, even when a single feature is manipulated while holding other correlated ones constant, it remains unclear how well this feature predicts impressions in real life when people are exposed to variation in facial features across many dimensions. It is possible that certain facial features are significantly related to personality impressions in highly controlled settings, but they might be poor predictors under more realistic conditions. For example, fWHR may be related to personality impressions when targets' gender, race, and approximate age are kept constant (as is often the case in social perception studies), but fWHR might not be an important cue when faces vary along many dimensions that are relevant for personality judgments. This limitation is exacerbated in studies using a two-alternative forced-choice design (Ormiston et al., 2017; Stirrat & Perrett, 2010). In this common experimental design, a face is manipulated on one dimension and two face versions are displayed side-by-side (e.g., high vs. low fWHR versions).

Participants then choose the face that they perceive as scoring higher on the relevant trait. As this approach highlights even subtle differences in facial features, it can produce effects that would not be observed with more naturalistic designs (A. L. Jones & Jaeger, 2019).

To address these limitations, some studies have used data-driven approaches, in which a large number of low-level facial characteristics (e.g., distances between different points in the face) are used to predict personality impressions (McCurrie et al., 2017; Oosterhof & Todorov, 2008; Song et al., 2017; Vernon et al., 2014). These techniques have proven very useful, for example, for visualizing prototypical configurations of faces (i.e., what a typical trustworthy-looking face looks like). However, because of their data-driven nature, it is often unclear to what extent results support theoretical predictions about the importance of different facial characteristics. For example, data-driven methods can be used to mathematically describe and visualize what a prototypically (un-)trustworthy face looks like (Dotsch & Todorov, 2012; Oosterhof & Todorov, 2008). Ratings of these prototypes might reveal that a trustworthy face scores higher on perceived femininity, babyfacedness, resemblance to a happy expression, and many other dimensions. Yet, this approach again provides limited insights into the *relative* importance of different psychological variables in impression formation.

Recent evidence also supports the predictive power of theory-driven variables. When comparing the predictive power of data-driven and theory-driven models for facial attractiveness, Holzleitner and colleagues (2019) found that the performance of a complex data-driven model was matched by using five theory-driven predictors at the same time, even though in isolation, these theory-driven predictors performed poorly. This speaks to the importance of identifying and testing theoretically important predictors at the same time, rather than in isolation, in order to build parsimonious and interpretable models of social perception.

The Current Study

In sum, previous approaches provide limited into which facial characteristics are central in impression formation. The current study was designed to address these limitations. We extend previous work in three crucial ways.

First, the majority of prior studies only examined one feature or one class of features in isolation (Said et al., 2009; Sofer et al., 2017; Stirrat & Perrett, 2010). Here, we test the relative importance of a wide range of features that are commonly studied in the literature. We examine four classes of predictors: resemblances to emotional expressions (e.g., resemblance to a happy expression; Said et al., 2009), demographic characteristic (e.g., gender and age; Sutherland et al., 2013), statistical characteristics (e.g., sex- and race-typicality Sofer, Dotsch, Wigboldus, & Todorov, 2015), and morphological characteristics (e.g., fWHR; Stirrat & Perrett, 2010). While theoretical approaches often focus on classes of facial features, grouping different features into a conceptually meaningful class is ultimately subjective. We therefore also examine the importance of all 31 facial features simultaneously.

Second, the majority of prior work has focused on the *explanatory* power of different facial features, testing how much variance in impressions is explained by different variables. However, this might overestimate the actual importance of specific characteristics due to overfitting (Yarkoni & Westfall, 2017). In the present study, we rely on machine learning to address this issue. For example, we use nested cross-validation and to compare the *predictive* power of different facial features (for similar applications of these methods, see Holzleitner et al., 2019; Jones & Jaeger, 2019).

Third, many prior studies were based on relatively small samples of stimuli (e.g., 50 or fewer; Carré et al., 2009; Sacco & Hugenberg, 2009; Stirrat & Perrett, 2012), which limits the generalizability of results. We therefore examine the predictors of personality impressions in a large and demographically diverse set of faces (n = 597). Our approach serves as a critical test of how well different characteristics—which have been theorized to be central for impression formation—predict personality impressions when faces vary (and are modeled) along a wide variety of different dimensions.

Methods

All data and analysis scripts are available at the Open Science Framework (https://osf.io/8rj7e/). We report how our sample size was determined, all data exclusions, and all measures.

Stimuli

We analyzed all 597 face images from the Chicago Face Database (Ma et al., 2015). All individuals wore a grey shirt, displayed a neutral facial expression, and were photographed from a fixed distance against a uniform background. The databases provides several advantages for the purpose of the current study. First, the database contains photographs of a large and diverse set of individuals who vary on gender (51.42% female), age (M = 28.86, SD = 6.30, Min = 16.94, *Max* = 56.38), and race (33.00% Black, 30.65% White, 18.26% Asian, 18.09% Latino). Thus, the image set represents a wide range of facial characteristics that people are exposed to in real life.

Variables

The database also contains a large number of objectively measured and subjectively rated characteristics for each face. Our aim was to predict average judgments of trustworthiness and dominance, which were rated on a 7-point scale. We examined the predictive power of 31 facial features, which we grouped into four classes of predictors.

Emotion resemblances included six variables representing the perceived resemblance of facial features to emotional expressions (anger, disgust, fear, happiness, sadness, and surprise).

Demographic characteristics included four variables: gender (coded 0 for male and 1 for female), race (Asian, Black, Latino, or White, with White coded as the reference category), and age. We also included a quadratic effect for age.

Statistical characteristics included four variables representing the perceived typicality of the face. We included gender-typicality (i.e., facial femininity), race-typicality, age-typicality (i.e., babyfacedness), and unusualness (i.e., how much the person would stand out in a crowd). We did not include facial masculinity as it was strongly correlated with femininity, $r(595) = -0.952$, $p < .001$.

Morphological characteristics included 15 variables that were selected based on a review of the social perception literature (Ma et al., 2015): face length, face width at the cheeks, face width at the mouth, face shape (face width at the cheeks divided by face length), heartshapeness (face width at the cheeks divided by face width at the mouth), nose shape (nose width divided by nose length), lip fullness (distance between top and bottom edge of lips divided by face length), eye shape (eye height divided by eye width), eye size (eye height divided by face length), upper head length (forehead length divided by face length), cheekbone height (distance from check to chin divided by face length), cheekbone prominence (difference between face width at cheekbones and face width at mouth divided by face length), face roundness (face width at mouth divided by face length), facial width-to-height ratio (distance between the outer edges of the cheeks divided by the distance between the upper lip and brow), and median luminance of the face. Even though it is not a morphological feature, we included luminance in this group of variables, as it constitutes another objectively measured, low-level stimulus property that has been linked to personality impressions (Dotsch & Todorov, 2012; Todorov et al., 2015).

Data on gender and race were directly provided by the photographed targets and morphological features were measured in Adobe Photoshop (Ma et al., 2015). All other characteristics represent mean ratings that were obtained from an average of 44 independent raters (*Min* = 21, *Max* = 131). All continuous predictors (except age) were *z*-standardized prior to analysis. A more detailed description of the variables and how they were measured is provided by Ma and colleagues (2015).

Analytic strategy

We use techniques borrowed from machine learning to estimate the predictive power of different (classes of) facial characteristics. For each model, we compute the root-mean-square error (RMSE), which represents the square root of the mean squared differences between predicted and observed values. In contrast to other statistics, such as R^2, RMSE has the advantage that it is not inflated by the number of predictors. Lower RMSE values indicate better predictive accuracy. We also computed adjusted R^2 for each model. Applying a penalty to the R^2 metric in line with the number of predictors in a model prevents, for example, that the morphology model outperforms the other models simply because it includes more predictors. We rely on cross validation—using the caret package (Kuhn, 2008) in R (R Core Team, 2020)—to avoid the problem of overfitting, in which a model is optimized to fit a particular data set to such an extent that it does poorly in predicting novel data (Yarkoni & Westfall, 2017).

Next to comparing different classes of facial characteristics, we also compare their unique predictive power by simultaneously entering all 31 characteristics into one regression model. Given that there were many significant correlations between cues (see Figure S1 in the Supplemental Materials), linear models may result in overfitted and highly variable estimates of the true importance of the parameters. To prevent this, we relied on Elastic Net regression (Hastie et al., 2009). Elastic Nets are linear models that simultaneously (a) shrink predictors to reduce overfitting through regularization and (b) perform variable selection by setting the coefficients of uninformative parameters to zero. Thus, this approach is ideally suited to examine the relative importance of different facial characteristics in predicting personality impressions.

Note that we do not report any inferential statistics. Indicators of statistical significance, such as *p*-values, depend on the size of a sample. As our analyses are based on sample sizes that can be arbitrarily increased, any difference would eventually be significant. Thus, a focus on comparing and interpreting effect sizes (e.g., the predictive accuracies of our models) is more appropriate (Troitzsch, 2014; see also Yarkoni & Westfall, 2017).

Results

Relative importance of classes of facial features

First, we compared the performance of different classes of facial characteristics in predicting perceptions of trustworthiness and dominance. To this end, we estimated linear regression models, in which trustworthiness ratings or dominance ratings were regressed on four classes of predictors (in separate models). We used the caret package (Kuhn, 2008) in R (R Core Team, 2020) to implement 10-fold cross validation with 100 repeats (i.e., 1000 resamples) to estimate the predictive accuracy of each model. The *emotions model* included six predictors representing resemblances to emotional expressions (e.g., resemblance to an angry facial expression). The *statistics model* included four predictors representing statistical properties of the faces (e.g., gender-typicality). The *demographics model* included four predictors representing demographic characteristics of the faces (e.g., gender). Finally, the *morphology model* included 15 predictors representing shape and color properties of the faces (e.g., facial width-to-height ratio). Detailed results for each model are presented in the Supplemental Materials.

For perceptions of trustworthiness (see Figure 1), the emotions model showed the best predictive accuracy (M_{RMSE} = 0.285, SD_{RMSE} = 0.026), followed by the statistics model (M_{RMSE} = 0.354, SD_{RMSE} = 0.030), the demographics model (M_{RMSE} = 0.388, SD_{RMSE} = 0.029), and the morphology model (M_{RMSE} = 0.402, SD_{RMSE} = 0.030). The same pattern was obtained when comparing how much variance was explained by the four models. The emotions model explained most variance (M_{R^2} = 0.531, SD_{R^2} = 0.083), followed by the statistics model (M_{R^2} = 0.274, SD_{R^2} = 0.090), the demographics model (M_{R^2} = 0.132, SD_{R^2} = 0.073), and the morphology model (M_{R^2} = 0.071, SD_{R^2} = 0.054).

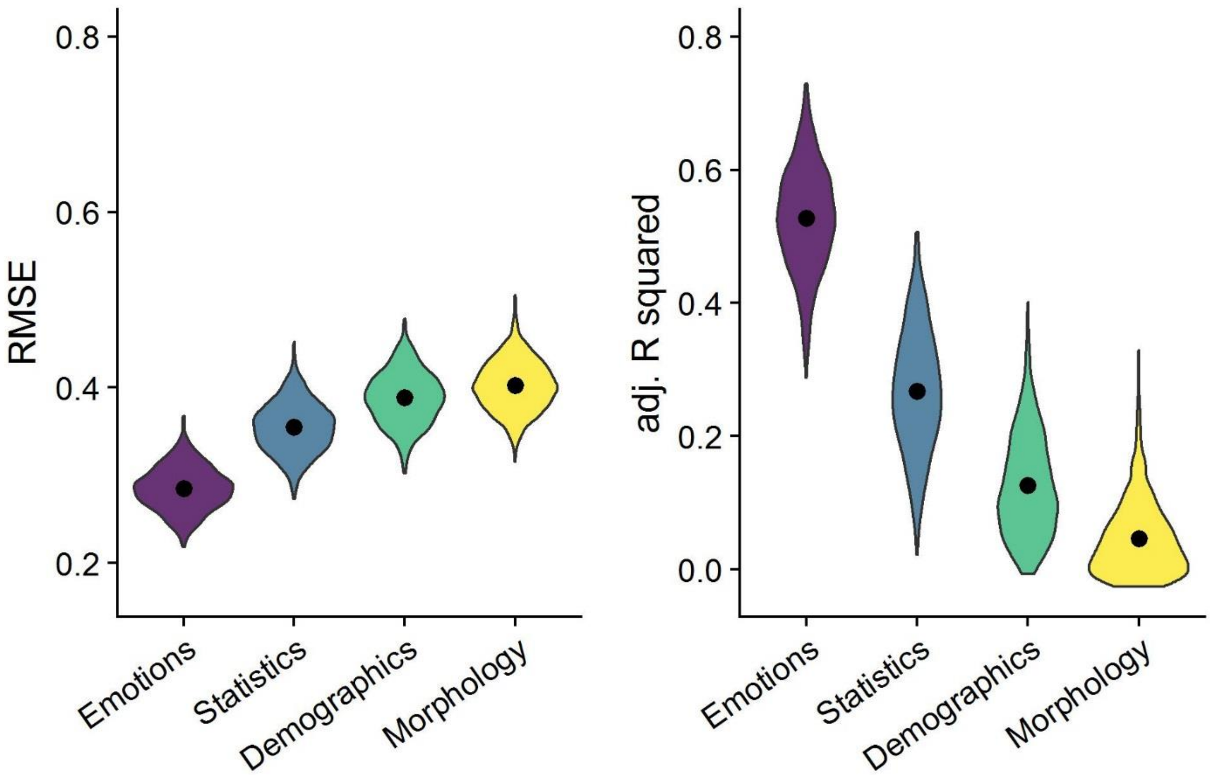

Figure 1. Violin plots showing performance of the four models in predicting trustworthiness impressions. Dots indicate the mean RMSE (left) and adjusted R^2 (right) from 10-fold cross validation with 100 repeats.

For perceptions of dominance, results were similar, albeit less pronounced (see Figure 2). The emotions model showed the best predictive accuracy (M_{RMSE} = 0.515, SD_{RMSE} = 0.042), followed by the statistics model (M_{RMSE} = 0.529, SD_{RMSE} = 0.041), the demographics model (M_{RMSE} = 0.535, SD_{RMSE} = 0.044), and the morphology model (M_{RMSE} = 0.574, SD_{RMSE} = 0.047). The same pattern was obtained when comparing how much variance was explained by the four models. The emotions model explained most variance (M_{R^2} = 0.423, SD_{R^2} = 0.092), followed by the statistics model (M_{R^2} = 0.391, SD_{R^2} = 0.087), the demographics model (M_{R^2} = 0.377, SD_{R^2} = 0.093), and the morphology model (M_{R^2} = 0.283, SD_{R^2} = 0.091).

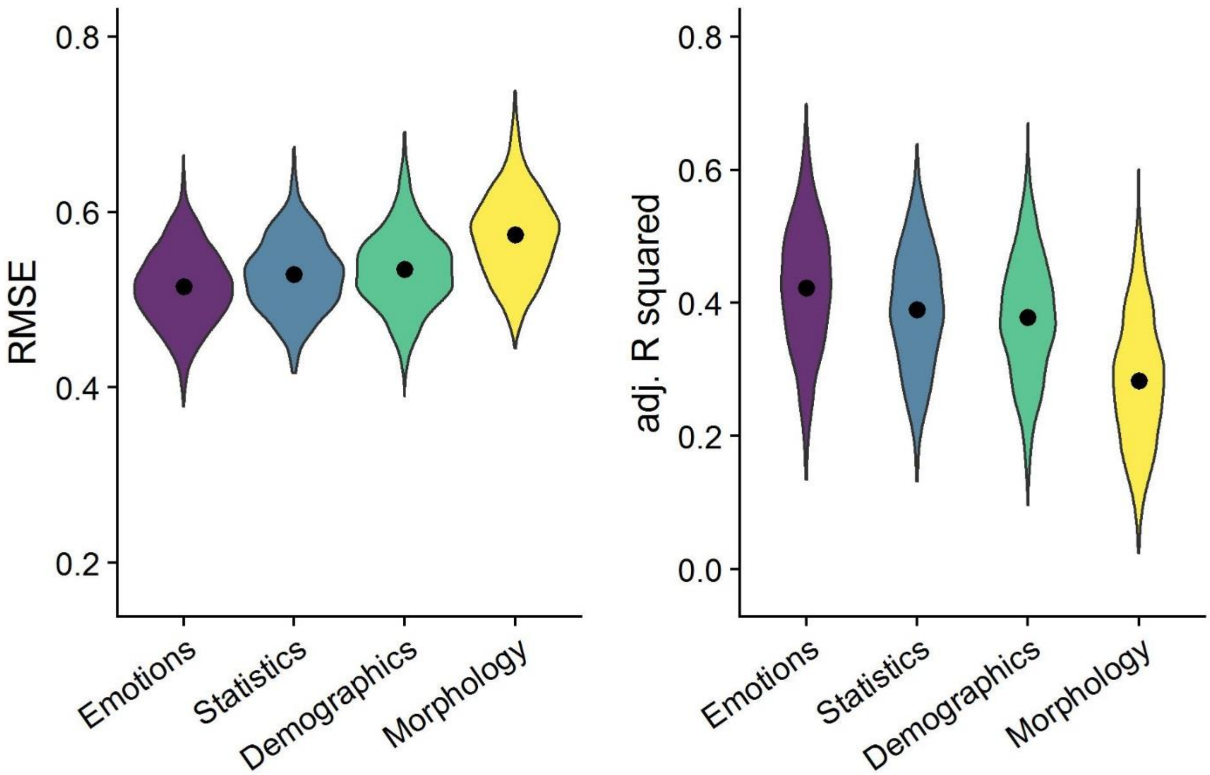

Figure 2. Violin plots showing performance of the four models in predicting dominance impressions. Dots indicate the mean RMSE (left) and adjusted R^2 (right) from 10-fold cross validation with 100 repeats.

Relative importance of all facial features

In the previous section, we examined the relative predictive power of different classes of facial characteristics. However, human perception of these classes of parameters occurs simultaneously, and observers may rely on some cues from a specific class of characteristics, while ignoring others. We therefore examined the influence of all 31 facial characteristics by simultaneously entering them into one regression model. We relied on Elastic Net regression (Hastie et al., 2009), which simultaneously (a) shrinks predictors to reduce overfitting through regularization and (b) performs variable selection by setting the coefficients of uninformative parameters to zero. The model has two hyperparameter that require tuning: *alpha*, which controls the degree of shrinkage, and the *L1 ratio,* which determines how aggressively coefficients can be set to zero. We additionally relied on nested cross-validation to ensure the robustness of our models. This involved splitting the full dataset into five folds. For each split of the data, a further 5-fold grid search cross-validation was carried out to derive the best hyperparameters before predicting the held out fifth fold. We repeated this process 20 times, yielding 100 training and test scores, and 100 sets of regression coefficients, which were averaged.

For trustworthiness, this approach revealed good performance for training and test RMSE (M_{train} = 0.583, SD_{train} = 0.014, M_{test} = 0.610, SD_{test} = 0.040) and for training and test R^2 (M_{train} = 0.659, SD_{train} = 0.018, M_{test} = 0.622, SD_{test} = 0.050), showing that our model was not overfitting the data. Examining performance on the test data showed that the model predicted trustworthiness impressions to within 0.61 points on a 7-point scale and explained 62.2% of the variance. Results revealed that the five most important predictors for trustworthiness impressions were resemblance to a happy facial expression, $\bar{\beta}$ = 0.412, facial femininity, $\bar{\beta}$ = 0.239, resemblance to an angry facial expression, $\bar{\beta}$ = -0.196, an unusual appearance, $\bar{\beta}$ = -0.128, and babyfacedness, $\bar{\beta}$ = 0.127 (see Figure 3).

Performance was similar for the dominance model, for training and test RMSE (M_{train} = 0.550, SD_{train} = 0.013, M_{test} = 0.578, SD_{test} = 0.038) and for training and test R^2 (M_{train} = 0.698, SD_{train} = 0.014, M_{test} = 0.659, SD_{test} = 0.045). Examining performance on the test data showed that the model predicted dominance impressions to within 0.58 points on a 7-point scale and explained 65.9% of the variance. The most important predictors were resemblance to an angry facial expression, $\bar{\beta}$ = 0.589, facial femininity, $\bar{\beta}$ = -0.252, resemblance to a sad expression, $\bar{\beta}$ = -0.172, being Asian (versus White), $\bar{\beta}$ = -0.162, and having lower facial luminance, $\bar{\beta}$ = -0.122 (see Figure 4).

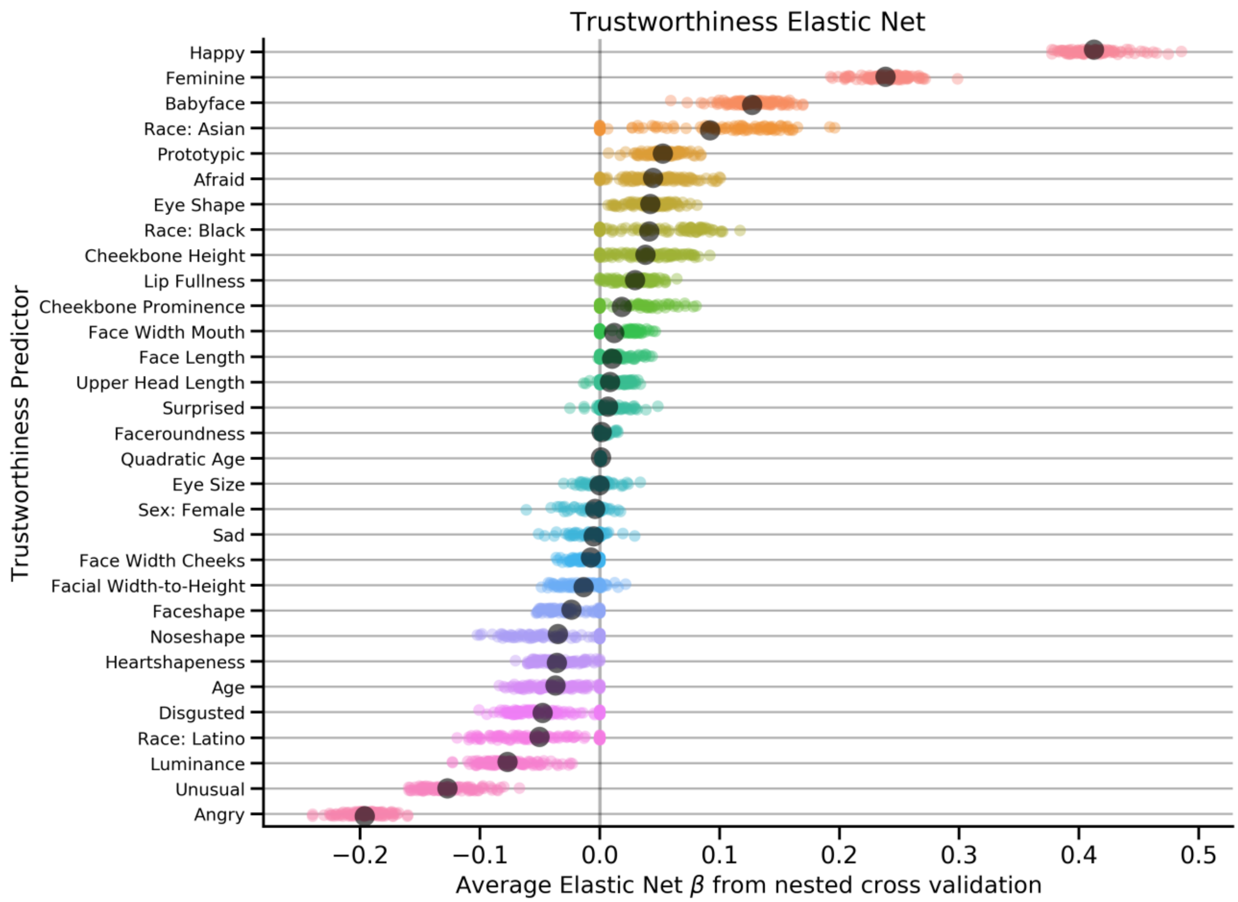

Figure 3. The relationship between different facial characteristics and trustworthiness impressions. Coefficients were derived from Elastic Net models with nested cross-validation.

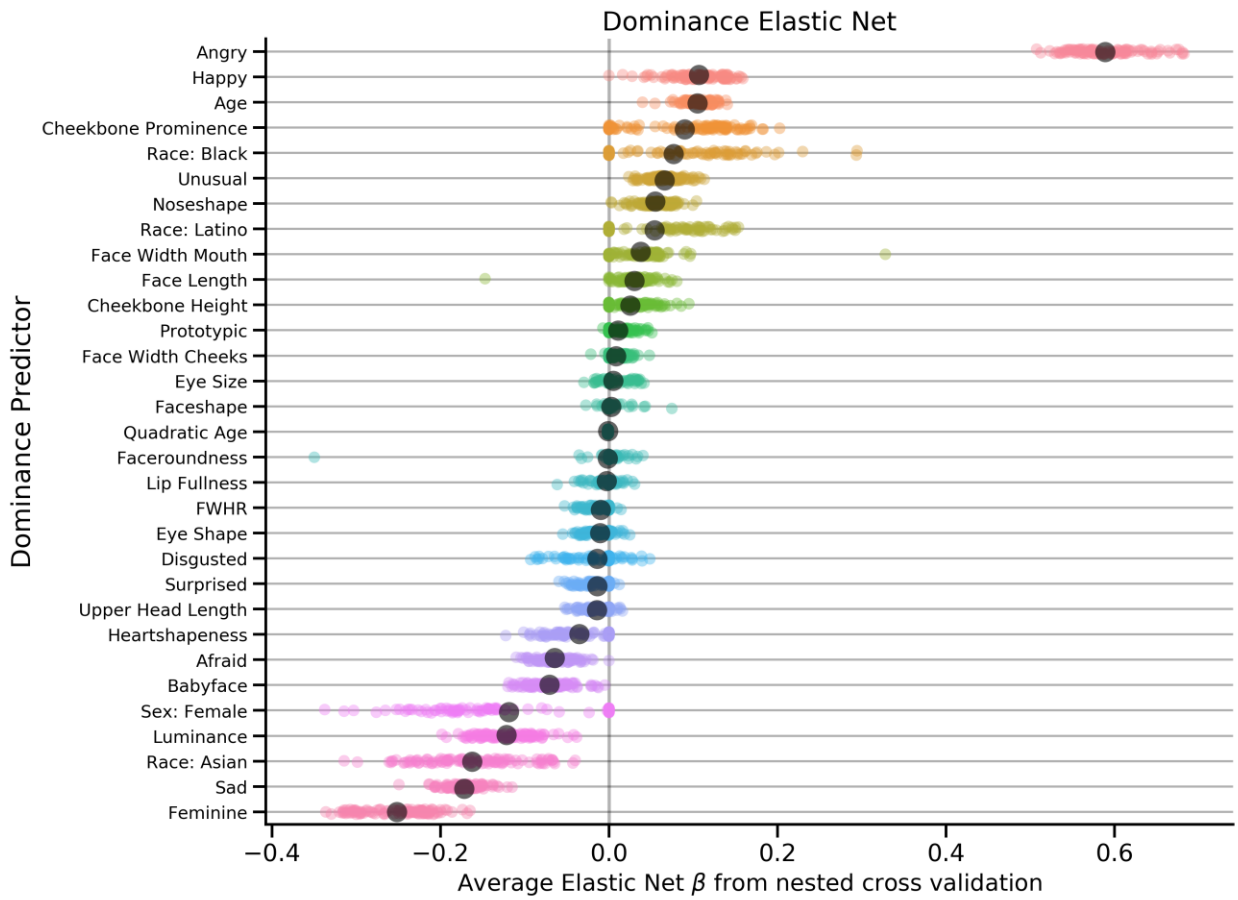

Figure 4. The relationship between different facial characteristics and dominance impressions. Coefficients were derived from Elastic Net models with nested cross-validation.

General Discussion

Which characteristics do people rely on to form personality impressions from faces? Some features, such as resemblances to emotional expressions and facial width-to-height ratio, occupy a central role in prominent theories of social perception (Todorov et al., 2008; Zebrowitz, 2017). However, it is not clear whether this focus is justified. Little is known about the relative importance of different characteristics because prior work has mostly examined one feature or one class of features in isolation. In short, it is not clear which facial features are actually central in impression formation. Here, we used machine learning methods to compare the extent to which a wide range of theoretically important facial features predict trustworthiness and dominance impressions for a large and demographically diverse set of faces.

Results showed that emotion resemblances (e.g., resemblance to a happy facial expression) were more predictive of trustworthiness and dominance impressions than statistical (e.g., sex-typicality), demographic (e.g., sex), or morphological features (e.g., facial width-to-height ratio). Moreover, examining the importance of all 31 facial characteristics showed that perceptions of trustworthiness were best predicted by a face's resemblance to a happy expression, whereas perceptions of dominance were best predicted by a face's resemblance to an angry expression. Thus, our results support the notion that resemblances to emotional expressions are central for explaining how people form personality impressions from facial features. This is in line with overgeneralization theory (Todorov et al., 2008; Zebrowitz, 2017), which posits that personality impressions of faces are driven by an oversensitive emotion detection system: Due to their social relevance, people even perceive emotions (and associated personality traits) in emotionally neutral faces that structurally resemble emotional expressions.

We also found that demographic factors (i.e., gender, age, and race)—which have received less attention as predictors of personality impressions—were often non-zero coefficients and, in some instances, among the top five predictors. This highlights potential problems associated with keeping features like gender and race constant when studying the basis of personality impressions. Certain features may have a significant influence when there is no variance in demographic characteristics, but they may be uninformative when demographics characteristics vary (as they tend to do in real life). For instance, a wealth of studies has examined the influence of facial width-to-height ratio on personality judgments (Ormiston et al., 2017; Stirrat & Perrett, 2010, 2012; Valentine et al., 2014). The current results show that, when modeled alongside a wide range of other features, fWHR was among the least informative predictors (with an average coefficient of -0.01 for the trustworthiness and dominance models). These findings suggest that the importance of fWHR as a basis for personality impressions may have been overstated in previous studies.

In general, low-level morphological features were less important than higher-level characteristics. While this finding may not be surprising, as high-level characteristics can by definition capture a more holistic and psychologically rich profile of a face, it is also not obvious. Prior work in social perception has shown that not all high-level variables that have been theorized to be psychologically meaningful actually turn out to be important predictors of people's judgments (Holzleitner et al., 2019). In fact, we find that some low-level characteristics (e.g., luminance, cheekbone prominence) were more important predictors of personality impressions than some high-level predictors (e.g., resemblance to an expression of disgust).

These results show that approaches as the current one are necessary to test whether theoretically important features are actually meaningful predictors of people's impressions.

Interestingly, all four classes of predictors that were examined here showed better predictive accuracy for trustworthiness perceptions than for dominance perceptions. For instance, the emotion resemblances model predicted trustworthiness impressions to within 0.29 points on a 7-point scale, whereas dominance impressions were only predicted to within 0.52 points. In a similar vein, the morphology model predicted trustworthiness impressions to within 0.40 points, whereas dominance impressions were only predicted to within 0.57 points. These results speak to the relative importance of different facial characteristics as determinants of trustworthiness and dominance impressions. Previous studies suggest that emotion resemblances are particularly important for trustworthiness impressions, whereas morphological characteristics, such as fWHR, are particularly important for dominance impressions (Hehman et al., 2015). However, the current results suggest that emotion resemblances are the most important determinant of both trustworthiness and dominance. It should also be noted that even though emotion resemblances were the most important class of predictors, not all emotion resemblances were equally meaningful. Resemblance to a happy expression was the most important predictor of trustworthiness impressions, whereas resemblance to an angry expression was the most important predictor of dominance impressions.

Limitations and Future Directions

Despite the relatively good performance of some of our models, results also suggest that our list of relevant features was not exhaustive. Emotion resemblances explained 53% and 42% of the variance in trustworthiness and dominance perceptions. Even the optimized Elastic Net models explained only between 60% and 66% of the variance, indicating there are other factors contributing to personality impressions. Candidate predictors may include facial characteristics that were not modeled here, but future studies could also investigate characteristics of the perceiver which explain a non-trivial amount of variance in impressions (Hehman et al., 2019). Examining the role of additional predictors will also show how generalizable our results are, as the relative importance of facial features may ultimately depend on the specific set of features that is modeled. Moreover, while the current set of faces was relatively large and diverse in terms of gender, age, and race, we only examined U.S. American individuals that were photographed in a controlled lab setting. Future studies could test whether the current findings replicate when using more naturalistic images of individuals from different nationalities (Sutherland et al., 2013).

References

Adams, R. B., Nelson, A. J., Soto, J. A., Hess, U., & Kleck, R. E. (2012). Emotion in the neutral face: A mechanism for impression formation? *Cognition & Emotion*, *26*(3), 431–441. https://doi.org/10.1080/02699931.2012.666502

Bijlstra, G., Holland, R. W., Dotsch, R., Hugenberg, K., & Wigboldus, D. H. J. (2014). Stereotype associations and emotion recognition. *Personality and Social Psychology Bulletin*, *40*(5), 567–577. https://doi.org/10.1177/0146167213520458

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

Deska, J. C., Lloyd, E. P., & Hugenberg, K. (2018). The face of fear and anger: Facial width-to-height ratio biases recognition of angry and fearful expressions. *Emotion*, *18*(3), 453–464. https://doi.org/10.1037/emo0000328

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.

Hehman, E., Flake, J. K., & Freeman, J. B. (2015). Static and dynamic facial cues differentially affect the consistency of social evaluations. *Personality and Social Psychology Bulletin*, *41*(8), 1123–1134. https://doi.org/10.1177/0146167215591495

Hehman, E., Stolier, R. M., Freeman, J. B., Flake, J. K., & Xie, S. Y. (2019). Toward a comprehensive model of face impressions: What we know, what we do not, and paths forward. *Social and Personality Psychology Compass*, *13*(2), 1–16. https://doi.org/10.1111/spc3.12431

Holzleitner, I. J., Lee, A. L., Hahn, A. C., Kandrik, M., Bovet, J., Renoult, J. P., Simmons, D., Garrod, O., Debruine, L. M., & Jones, B. C. (2019). Comparing theory-driven and data-driven attractiveness models using images of real women's faces. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(12), 1589–1595. https://doi.org/10.1037/xhp0000685

Jones, A. L. (2019). *Beyond average: Using face regression to study social perception*. https://doi.org/10.31234/osf.io/dpmzq

Jones, A. L., & Jaeger, B. (2019). Biological bases of beauty revisited: The effect of symmetry, averageness, and sexual dimorphism on female facial attractiveness. *Symmetry*, *11*(2). https://doi.org/10.3390/sym11020279

Jones, B. C., DeBruine, L. M., Flake, J. K., Aczel, B., Adamkovic, M., Alaei, R., Alper, S., Álvarez Solas, S., Andreychik, M. R., Ansari, D., Arnal, J. D., Babincák, P., Balas, B., Baník, G., Barzykowski, K., Baskin, E., Batres, C., Beaudry, J. L., Blake, K. R., … Chartier, C. R. (2019). *To which world regions does the valence-dominance model of social perception apply?* https://psyarxiv.com/n26dy/

Kosinski, M. (2017). Facial width does not predict self-reported behavioral tendencies. *Psychological Science*, *28*(11), 1675–1682. https://doi.org/10.1177/0956797617716929

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 159–160. https://doi.org/10.1053/j.sodo.2009.03.002

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

McCurrie, M., Beletti, F., Parzianello, L., Westendorp, A., Anthony, S., & Scheirer, W. J. (2017). Predicting first impressions with deep learning. *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*, 518–525. https://doi.org/10.1109/FG.2017.147

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

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

Ormiston, M. E., Wong, E. M., & Haselhuhn, M. P. (2017). Facial-width-to-height ratio predicts perceptions of integrity in males. *Personality and Individual Differences*, *105*, 40–42. https://doi.org/10.1016/j.paid.2016.09.017

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

Sacco, D. F., & Hugenberg, K. (2009). The look of fear and anger: Facial maturity modulates recognition of fearful and angry expressions. *Emotion*, *9*(1), 39–49. https://doi.org/10.1037/a0014081

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, *9*(2), 260–264. https://doi.org/10.1037/a0014681

Sofer, C., Dotsch, R., Oikawa, M., Oikawa, H., Wigboldus, D. H. J., & Todorov, A. (2017). For your local eyes only: Culture-specific face typicality influences perceptions of trustworthiness. *Perception*, *46*(8), 914–928. https://doi.org/10.1177/0301006617691786

Sofer, C., Dotsch, R., Wigboldus, D. H. J., & Todorov, A. (2015). What is typical is good: the influence of face typicality on perceived trustworthiness. *Psychological Science*, *26*(1), 39–47. https://doi.org/10.1177/0956797614554955

Song, A., Li, L., Atalla, C., & Cottrell, G. (2017). *Learning to see people like people*. http://arxiv.org/abs/1705.04282

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*(3), 349–354. https://doi.org/10.1177/0956797610362647

Stirrat, M., & Perrett, D. I. (2012). Face structure predicts cooperation: Men with wider faces are more generous to their in-group when out-group competition is salient. *Psychological Science*, *23*(7), 718–722. https://doi.org/10.1177/0956797611435133

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

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Troitzsch, K. (2014). Analysing simulation results statistically: Does significance matter? In D. Francisca, G. Pereira Dimuro, & H. Coelho (Eds.), *Interdisciplinary Applications of Agent-Based Social Simulation and Modeling* (pp. 88–105). IGI Global.

Valentine, K. A., Li, N. P., Penke, L., & Perrett, D. I. (2014). Judging a man by the width of his face: The role of facial ratios and dominance in mate choice at speed-dating events. *Psychological Science*, *25*(3), 806–811. https://doi.org/10.1177/0956797613511823

Vernon, R. J. W., Sutherland, C. A. M., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, *111*(32), E3353–E3361. https://doi.org/10.1073/pnas.1409860111

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science*, *26*(3), 237–242. https://doi.org/10.1177/0963721416683996