Overgeneralizing Emotions:

Facial Width-To-Height Revisited

Sabine Windmann, Lisa Steinbrück, and Patrick Stier

Johann-Wolfgang-Goethe University Frankfurt/Main, Germany

Author Note

Author contributions

All authors developed the study concept together, and all contributed to the study design. L.S. and P.S. performed data acquisition and preprocessing in Study 1 supervised by S.W. who also coordinated and supervised data acquisition in Studies 1b, 2, and 3. Data analysis in Study 1 was prepared by L.S. and P.S., and was carried out and extended to Studies 2 and 3 by S.W. Manuscript and revisions were written by S.W. with L.S. and P.S. providing editorial assistance. All authors checked and approved the final version of the manuscript.

Abstract

Facial width-to height ratio (fWHR), presumed to be shaped by testosterone during puberty, has been linked with aggressive, dominant, and power-seeking behavioral traits in adult males, although the causal mediation is still being disputed. To investigate the role of mere observer attribution bias in the association, we instructed participants to draw, feature-assemble, or photo-edit faces of fictitious males with aggressive-dominant character (compared to peaceloving-submissive), or powerful social status (compared to powerless). Across three studies involving 1,100 modeled faces in total, we observed little evidence for attribution bias with regards to facial width. Only in the photo-edited faces did character condition seem to affect fWHR; this difference, however, relied on displayed state emotions, not on static facial features. Anger, in particular, was expressed by lowered or V-shaped eyebrows, whereby facial height was reduced so that fWHR increased, relative to the comparison condition where the opposite happened. Using Bayesian analyses and equivalence testing, we confirmed that, in the absence of state emotionality, there was no effect of character condition on facial width. Our results add to a number of recent studies stressing the role of emotion overgeneralization in the association of fWHR with personality traits, an attributional bias that may give rise to a self-fulfilling prophecy. Methodologically, we infer that static images may be of limited use for investigations of fWHR because they cannot sufficiently differentiate between transient muscular activation and identity-related bone structures.

Keywords: facial features, fWHR, anger, aggression, attribution bias.

Revisiting Facial Width-To-Height

Modern theories conceptualize the apprehension of emotional and motivational social cues as part of a perception-action cycle (Chatterjee & Vartanian, 2016; Elshout et al., 2017; Fuster, 2006). To guide adaptive behavior, predictions are being made on the basis of prior experiences, semantic knowledge, and perceptual imagery, at both intuitive and strategical levels of processing, and are being updated continuously by (consciously or unconsciously) perceived environmental feedback.

Humans have been shown to infer state and trait information from observed facial feature configurations within fractions of a second (Bar et al., 2006; Willis & Todorov, 2006). Rapid and automatic as these inferences are, they are also subject to biases and overgeneralizations (Todorov et al., 2015; Zebrowitz, 2017; Zebrowitz & Montepare, 2008), especially when threat cues are implicated (LeDoux & Daw, 2018). Due to their high survival value, threat cues in social and other contexts are predestined to bias perceptions and memories (Windmann & Krüger, 1998).

Perhaps the most extensively researched threat-related facial feature is the facial width-to-height ratio (fWHR). It refers to a size-independent measure of facial width taken from 2D photographs, namely the bizygomatic distance (between left and right cheekbones) divided by facial height, taken usually from the lower bound of the nasial edge of the eyebrows down to the cupid's bow (i.e., the upper bound of the upper lip; Carré et al., 2009; Geniole et al., 2015). The feature has originally been described as sexually dimorphic (Weston et al., 2007), and taken as a proxy for testosterone exposure during puberty (Lefevre et al., 2013).

Carré, McCormick and colleagues were the first to investigate the psychological implications of fWHR in demonstrating that white males with higher fWHR behave more aggressively in behavioral assessments both inside and outside laboratory contexts (Carré &

McCormick, 2008). They later showed that observers estimate a male target's trait aggression from fWHR after viewing facial images for only 39 ms (Carré et al., 2009).

Over the following years, an impressive amount of concordant evidence accumulated. White males with high fWHR were found to be more aggressive and dominant in both self-reports and behavioral lab assessments compared to men with low fWHR (Geniole et al., 2014; Lefevre et al., 2014; Lefevre & Lewis, 2014; MacDonell et al., 2018). Specifically, they were found to score higher on trait measures of psychopathy (Anderl et al., 2016; Geniole et al., 2014; Noser et al., 2018), to be more likely to cheat or exploit others in experimental game settings (Geniole et al., 2014; Haselhuhn & Wong, 2012; Stirrat & Perrett, 2010), and to more readily endorse racial prejudices (Hehman et al., 2013). Field studies added ecological validity to these findings by showing fWHR to correlate with achievement drive in US American presidents (Lewis et al., 2012) and Chinese financial analysts (He et al., 2019), with game performance of and/or penalties received by professional hockey, baseball, football players, and martial arts fighters (Carré & McCormick, 2008; Trebicky et al., 2015; Tsujimura & Banissy, 2013; Welker et al., 2015), and with leadership status and/or leadership performance in large companies (Ahmed et al., 2019; Alrajih & Ward, 2014; Hahn et al., 2017; Kamiya et al., 2019; Mills & Hogan, 2020). Taken together, and summarized by two independent meta-analyses (Geniole et al., 2015; Haselhuhn et al., 2015), fWHR has been linked with aggression, dominance, and social status in a variety of research settings.

It is only more recently that some studies have shed doubt on this conclusion. For example, Kosinski (2017) did not find any significant relation between fWHR and 55 well-established psychometric constructs, including impulsiveness and sense of fairness, in a large dataset of more than 137,000 individuals. Further, Wang and colleagues (2019) did not find any significant association between fWHR and self- or other-reported antisocial behavioral

tendencies in $N = 1,305$ business executives, and attribute previously published effects to biased perceptions. Likewise, no association was found between fWHR and aggressive behavior in more than 400 professional football players (Krenn & Meier, 2018). Finally, the hypothesized mediating effects of testosterone on fWHR have been questioned based on recent findings (Bird et al., 2016; Eisenbruch et al., 2018; Kordsmeyer et al., 2019).

On the one hand, the inconsistencies in the literature may be related to altered research and publication practices in the aftermath of the replication crisis (c.f., Kosinsky, 2017). On the other hand, they may point to methodological issues and mediating factors (other than testosterone) which can sometimes simulate a positive correlation. For example, when chief executive officers (CEOs) are found to have wider-than-normal faces on portrait photographs published on the internet (c.f., Hahn et al., 2017; Mills & Hogan, 2020), this provides no direct evidence that these individuals are indeed more aggressive and dominant. Instead, superior camera gear and settings may have been used in their photography (Trebicky et al., 2016), or the to-be-published portraits may have been selected or edited in a particular way, perhaps to match the public image of a strong leader (see discussion in Hahn et al., 2017). Likewise, when politicians with high relative to low fWHR are found more likely to be convicted of corruption, and study participants indeed categorize these individuals more often as corrupt-looking (Lin et al., 2018), this does not necessarily mean that corruptibility can be inferred from the width of the politicians' faces. It may as well indicate that if a politician's fWHR triggers the mere attribution of dishonesty, he/she might be offered bribes more often; or it could mean that politicians with wider faces are more often suspected of, investigated for, and/or found guilty of corruption. Hence, outside proper experimental control, a positive link between fWHR and certain behavioral traits can be mimicked by attributional biases, whether or not the link truly exists.

Alternatively, and particularly relevant from the perspective of a social perception-

action cycle, even if the association between fWHR and antisocial traits truly exists, it may have evolved only due to attributional biases. A number of studies show that observers do indeed perceive and interpret fWHR as a signal of threat, dominance, and lowered trustworthiness (Geniole et al., 2015; Haselhuhn et al., 2013; Matsumoto & Hwang, 2021; Ormiston et al., 2017; Stirrat & Perrett, 2010; Valentine et al., 2014). They even adjust their behavior to the fWHR of the target individual, e.g., by altering cooperation rates (Haselhuhn et al., 2013; Stirrat & Perrett, 2010), approach behavior (Lieberz et al., 2017), and short-term oriented mate choice (Valentine et al., 2014). Hence, to the degree that observers believe in fWHR as a signal of threat, they might treat men with wider faces less welcoming than others. This would in turn trigger skeptical or otherwise adverse reactions in target individuals which they feed back to observers, paving the way for a vicious cycle (Elshout et al., 2017). Haselhuhn and colleagues have sketched this unwholesome dynamic empirically (Haselhuhn et al., 2013). They found that men with wider faces get allocated fewer resources in a decomposed economic game by observers than narrow-faced males, and that the relatively low shares that these men typically receive are countered by further reductions in cooperation and prosociality. Similar dynamics can be expected to unfold when women show reduced approach behavior towards men with high compared to low fWHR (Lieberz et al., 2018; Lieberz et al., 2017), when wider-faced men are ascribed to more anger and less humanness (Deska et al., 2018a, 2018b), and when are more likely to be judged criminal liars (Matsumoto & Wang, 2021). Most dramatically, one study showed that wider-faced convicts received the death penalty more often than narrow-faced convicts (Wilson & Rule, 2015).

Needless to say, when people get treated differently for the shape of their faces continuously throughout their lifetimes or at significant stages of their development, they might adapt over time in terms of their social personality development. That is, they might wind up their habitual aggression and dominance to be able to counter the continuous

depreciative treatment, thereby establishing a self-fulfilling prophecy (Haselhuhn et al., 2013). This theorizing acknowledges fWHR as a valid cue for social threat, but one that is rooted in observers' attribution patterns – not in target's predispositions – owing to the feedback loop characterizing social interaction dynamics.

The present study did not ask whether the link between fWHR and antisocial behavioral traits is valid or not. Instead, we asked whether observers *believe* that it is. In other words, we aimed to identify the role of mere observer bias in the link of FHWR and antisocial behavioral traits, independent of the actual connection. Therefore, we isolated perceiver-driven ("mind"-driven) components from target-driven components (Xie et al., 2019). To this end, we cut in on the social perception-action cycle at the point where observers' beliefs about the behavioral implications of fWHR may come into play. Specifically, instead of presenting participants with selected pictures of male faces varying in fWHR and asking them to rate these men's presumed personality traits, as typically done in fWHR research, we presented participants with selected personality traits and asked them to produce the associated faces from their imagination using various (semi-)artistic techniques. We then took fWHR from the generated faces, and tested for differences between the character trait conditions. In so doing, we determined participants' beliefs about the association between fWHR and antisocial traits, as residing in their minds, regardless of the actual validity of the association in the outside world. In this sense, we "revisited" fWHR from the perspective of the observer.

Artistic image generation tasks

Representational visual arts production has long been used in Psychology to tackle the semantic memory and imagery components of mental representations in diagnostic (Gu et al., 2020), educational (Mukunda et al., 2019), developmental (Sayõl, 2001), and therapeutic contexts (Fish, 2019). Here, we used three artistic techniques which varied with regards to

amount and quality of the perceptual input provided, from entirely lacking (empty sheet), over sketchy and deranged (array of facial features), to realistic. The three tasks were a drawing task, an assembly task, and a digital photo-editing task, respectively.

Drawings have been used before in research investigating social stereotyping (Miall et al., 2009) and emotion conception (Sayõl, 2001; Terada & Takeuchi, 2017). They tap into mental imagery because they evolve in the absence of externally driven stimulus processing (Abraham, 2018; Dijkstra et al., 2019; Miall et al., 2009). The produced image, however, will not only reflect imagery skills, but also drawing skills, depending on how well the imagination can be translated into the appropriate motor output.

We therefore designed our second technique as an assembly-kit. In this task, we provided participants with an array of pre-designed sketches of head shapes and facial elements (eyes, eyebrows, noses, lips), out of which they selected the preferred ones to configure the face. In her work with children, Sayõl (2001) showed that such *selection-construction* is much easier than free drawing when participants attempt to express their understanding of facial feature configurations. Similarly, Finke (1990) and likewise Verstijnen et al. (1998) showed that *combining* geometrical shapes to construct a figure is performed equally well by experts and novices, unlike *restructuring* the materials, a more innovative creation process, which benefits from expertise.

Finally, to complement the two tasks providing absent (drawings) or misarranged/sketchy stimulus input (assembly-kit), we used a third method that provided realistic perceptual input. Specifically, we used the template of a male's face in near photo quality that needed to be edited using specific computer software. Due to the realistic perceptual input provided, in connection with the simple response requirements of the task, this method can be presumed to be less dependent on imagery and construction skills than the other two (drawing and assembly), and instead to rely more on bottom-up activated

processing (Dijkstra et al., 2019, Fan et al., 2020; Michelon & Koenig, 2002). Evidently, as the same template was shown in both, the experimental condition and the comparison condition, the perceptual input provided in this task was still entirely ambiguous with regards to the connection between fWHR and character trait.

Research Overview

In summary, we had participants generate facial images of "bad guys" versus "good guys" with three methods: free drawings, feature assembly, and photo-edit, and assessed fWHR in the generated images. Because we observed several faces to show emotional expressions, we conducted another study (Study 1b) to examine the influence of displayed state-emotionality on the experimental effects. Studies 2 and 3 replicated and extended these findings for the drawings and the photo-editing method, in trying to experimentally (in addition to statistically) control for the potentially confounding effects of state-emotionality.

Open Science

Open Science practices were followed throughout. All studies were preregistered. Sample sizes were determined beforehand based on power considerations. All data exclusions, all measures taken, and all manipulations performed have been reported. Many extra analyses were performed to test the robustness of the results, which led to extensive supplementary materials. Materials, scripts, and data of all studies are publicly available (Windmann, Steinbrück, & Stier, 2021). All study protocols have been approved prior to data collection by the internal ethics review board of our faculty.

Study 1a

In the first of our series of studies, we had participants express their ideas about the facial characteristics of an aggressive-dominant man in the three image production tasks (free drawing, assembly of features, and photo-editing). We expected to tap more into bottom-up driven perceptual and memory processes when a realistic template was provided (photo-edit)

compared to a sketchy (assembly) or entirely lacking perceptual input (free drawing) that would rely more on top-down imagination (Dijkstra et al., 2019; Fan et al., 2020; Michelon & Koenig, 2002).

In a full factorial, within-subject's design, all participants were asked to use all three methods sequentially to create the facial image of a man with an "aggressive-dominant" character in one condition versus a "peaceloving-submissive" character in the comparison condition. The comparison condition was not meant to yield a "neutral" comparison point but was deliberately chosen to maximize the contrast. For all three methods of image generation, participants were instructed to create the faces in the two conditions the way they believed the general public to imagine them. This was done to prevent participants from developing concerns over having their own prejudicing assessed. Participants were told further that the faces should belong to a fictional (i.e., non-existing) white male with neutral facial expression in frontal orientation.

After data collection, two raters blind to the research question and the research itself determined facial width-to-height ratio (fWHR) from all pictured faces, which were then analyzed for effects of character condition using ANOVA. The study has been preregistered at https://doi.org/10.17605/osf.io/3zqjc.

Methods

Participants

One hundred thirty-six students participated for course credit ($N = 94$ undergraduates of Psychology; of these, 48% in their 1st and 34% in their 3rd semester) or financial compensation (€ 10, ca. 60 minutes) and a small gift to choose (pen, college block, etc.). Three participants were excluded because they stated that they had modeled the face of a real man in at least one of the three tasks. No participant guessed the research question correctly upon questioning. The final sample consisted of 133 participants aged 18 to 56 years ($M =$

22.5, *SD* = 5.3); 110 female, 22 male, one diverse[1].

The required sample size had been determined beforehand using G-Power (Faul et al., 2007). With $N = 108$, a medium effect size of $f = .25$ is detected with a power of .95 and the standard .05 significance threshold by an ANOVA with a two-factorial design involving 3 groups x 2 uncorrelated measurements, which is less powerful than our actual design. As preregistered, we had anticipated that 10-20% of the data sets would have to be excluded from data analysis primarily due to problems with fWHR assessment (particularly for the drawings where eyebrows mere missing in some cases, contours were unclear etc). Hence, to account for the anticipated loss, we preregistered $N = 130$ participants, but ended up including six more who had already signed up for participation by the time we reached the preregistered number.

Materials

For the drawings, a sharpened pencil, two blank Din-A4 sheets of paper, and an eraser were used. For the assembly-kit, predesigned parts of a face (8 facial shapes, 7 pairs of eyes, 7 pairs of eyebrows, 7 lips and 7 noses, all of which varied only in size) were presented on the left side of a 22" computer screen in PowerPoint (Microsoft©, Redmond, Washington). The facial elements were locked so that they could be moved across the screen, but not modified, not even in size. Beneath two black vertical lines, the right of the screen provided an empty space for the to-be-assembled face.

For the digital editing of the facial photographs (photo-edit method), we used the commercially available FaceGen© Artist software by Singular Inversions Incorporated, Toronto, Canada. The program allows for manipulation of 61 facial features of a model face in near-photograph quality using sliders sorted into 10 facial regions. Importantly, when one slider is used, other features go along automatically so that the resulting face still looks

---

[1] "Diverse" denotes the official third gender category used in Germany.

realistic. The program contains three sliders for "cheekbones", and two for "cheeks", starting at position 5, directly below four sliders for "eyebrows". In both experimental conditions, and to all participants, the same frontally-oriented template of an average European male aged 30 was provided as starting point. Although presenting various different faces would have been preferable in terms of generalizability (Simons et al., 2017), we wanted to avoid the associated error variance in this first study of its kind, as in a proof-of-concept. Note, however, that the template did not represent a randomly chosen males' face but a statistical prototype generated by Singular Inversions using principal component analysis of 273 high-resolution 3D face scans, to some benefit of the external validity of the setup.

A digital stop watch, visible to participants on a second screen, was used to monitor the time. On that same screen, a number of standardized questions were presented by the survey tool EFS Survey© by Questback (Cologne, Germany) after the face modeling tasks. These questions were, first, three statements that we had created ourselves to determine participants´ individual tendency to prejudice based on overt bodily features (e.g., "The character of a person is mirrored in his/her face"); responses were taken on a 100 point visual analogue scale. Secondly, we asked: "Have you tried in any of the images you generated to depict a person who you know?" (yes/no); third, we provided open-formatted questions tapping into prior knowledge of fWHR research: "What do you think our experiment dealt with?" "Which facial features are we interested in?", and "What could be our hypothesis?". Finally, demographic variables were taken (age, gender, studied major and semester, period of time living in Germany).

Procedures

Data were collected in groups of three participants in the laboratory of our work group with shielding screens between them. Written informed consent was taken prior to participation.

Participants were informed that they would participate in a study on the impact of prejudices on the association between facial features and character traits, and would be asked to draw, model, and manipulate faces of male adults with a particular character. It was stressed that their personal prejudicing would not be relevant for the face modeling tasks since the instructions would only refer to the participant's view of commonly held stereotypes, not to their own stereotyping or prejudicing. Two experimenters (coauthors P.S. and L.S.) who were present during the entire data collection phase provided the instructions, both orally and in writing.

The three methods (drawing, assembly, photo-edit) were employed in block-randomized order (see supplementary materials S1). Across participants, the order of the two trait conditions (aggressive-dominant versus peaceloving-submissive) was also quasi-randomized, but held constant within a given group of participants and across all three methods to avoid any confusion.

Participants were given separate instructions prior to each method. For the drawings, they were asked to draw a fictitious man of western origin with neutral facial expression like in a biometric passport photograph. To decrease any pressure and potential embarrassment, participants were told that their artistic talent was not at stake, and that sketches were sufficient. They were also told that the drawn faces could be completely different from the faces generated with the other two methods.

For the assembly-kit method, participants were instructed to use the facial elements displayed on the left side of the computer screen to configure a face in accordance with the current character condition on the right side of the screen. We asked participants to model the proportions as realistically as possible.

For the photo-edit method, the program FaceGen© was launched by one of the experimenters to ensure correct loading of the template photograph. Participants were

explained the overall arrangement and were allowed to try out the rulers as much as they wanted. Subsequently, they were asked to model a realistically proportioned face with a neutral expression in both conditions, and were encouraged to use all rulers for that purpose.

Additional details on the procedures are given in the supplementary materials S1. In accord with pilot testing, participants were given three minutes for the assembly-kit, five minutes for the drawing, and 10 minutes for the photo-edit task. These times proved sufficient as we observed that many participants began to model irrelevant details of the faces towards the end of these periods. For all methods, participants were instructed to model realistic faces as much as possible with an emotionally neutral expression.

When the time period for the modeling of a face had elapsed, participants were asked on the left screen to indicate by mouse-click from 0 (not at all) to 100 (very much), how i) *satisfied* they were with their image [in German: "Wie gut ist es Ihnen gelungen, das Gesicht eines aggressiv-dominanten Mannes zu modelieren?"], ii) how *natural/realistic* they found the generated face [in German: "Wie natürlich/realitätsnah schätzen Sie den von Ihnen gezeichneten Mann ein?"], and iii) how attractive they found the generated face from a common viewpoint [in German: "Wie schätzen Sie die Attraktivität des von Ihnen erstellten Männergesichts aus Sicht der Allgemeinheit ein"?]. These three ratings (obtained in both trait conditions from each participant) will be referred to as *satisfaction, naturalness,* and *attractiveness* ratings.

After completion of the image generation tasks with all three methods, participants were given back their drawings, and were asked to mark with a fine-liner the position of the nasion and of the left and right cheekbones. Experimenters explained beforehand, by pointing to the relevant locations in their own faces, what the nasion and the cheekbones were and where they would be expected to be located approximately. Finally, participants were thanked, debriefed, and paid; they were given the opportunity to leave an email address to obtain further

information on the results of the study, if of interest, and were asked to refrain from conveying any information on the study goals to any potential future participants.

FWHR measurement

All images were converted to (or scanned into) .png files, and presented in quasi-randomized order (randomized within methods) to two raters (L.K. and A.F.) who were entirely naïve regarding our research questions and research goals. Using ImageJ (https://imagej.net), the two raters independently determined (in pixels) (A) facial width as the horizontal distance of the facial contours at the height of (rater-determined) cheekbones, (B) facial height as the distance between the upper bound of the upper lip (at the intercept with the vertical midline of the face) and the lower bound of the mid eyebrows, taking the lower one of the two if these were not aligned vertically. Division of (A) by (B) yields fWHR (Carré et al., 2009; Geniole et al., 2015) as used in the majority of prior studies. Because there is no prior experience with fWHR taken from artificial facial images, we were unsure which feature marks the raters could assess reliably in most cases, especially with regards to the drawings. Hence, we additionally had raters take alternative measures of facial width, depending on method (detailed below). Additionally, for all three methods, we had raters determine one alternative measure of facial height, namely the distance between the upper bound of the upper lip (covering the prosthion) and the nasion (Weston et al., 2007). This was done because the literature is not clear as to which of the height measures is to be preferred.

Raters were trained with independent materials. They were instructed to code meticulously any uncertainties they felt in defining the marks and taking the measures. The coding scheme was very detailed and individualized for each method. This was particularly relevant for the drawings. Here, raters were instructed to use participant's marks of cheekbones and nasion whenever these appeared plausible. If these were clearly misplaced, however, or severely misaligned, raters measured at sites that they found more realistic. Since

the raters were entirely naïve, and worked independently without communicating with each other, these corrections should be completely unbiased. In addition, analyses of fWHR based strictly on participants' marks, although adding error variance, provided strikingly similar results (Stier, 2019) and did not change our conclusions; all data are provided at https://doi.org/10.17605/osf.io/mtzax. Other problems were noted when i. the face was not frontally oriented (=> missing data), ii. the lip was drawn as only one line (=> raters used the upper bound of this line, i.e., the intercept of that line with the vertical midline of the face) iii. eyebrows were missing (=> missing data), iv. the uppermost edge of the upper lip was located laterally (=> raters used the center of the lip at the intersect with the vertical midline of the face), v. other problems (to be detailed in writing). Coding of multiple problems was allowed.

From the assembly-kits, only facial height had to be determined by the raters because facial width was taken from the templates. This presented no problems for the eyebrows. However, the nasion was often a problem with this method, typically because the selected nose was disproportionately small, hence the nasion was located too low vertically. In fact, the marked nasion was replaced by *both* raters 71 times in the aggressive-dominant condition and 63 times in the peaceloving-submissive condition, so in more than half of the images, and was therefore not used any further. Note, however, that this problem occurred only in the assembly kit method due to the selection of too small features, and only for the position of the nasion which was not needed for the main analyses anyway. Another problem code was provided for eyebrows that were not aligned vertically.

For the photo-edit task, facial width was taken, first, as the distance between the outer rim of the cheekbones, and secondly, as the distance between the leftmost and rightmost points of the facial contours at the height of the cheekbones. The former was used for the present analyses, but results did not change in any relevant way when the contour-based

measure was used (Stier, 2019); data are provided at https://doi.org/10.17605/osf.io/mtzax).

Raters were asked to code the following problems: i. cheekbones were indiscernible, ii.

nasion was indiscernible, iii. lips were asymmetric (=> raters used the intercept of the upper

bound of the lip with the vertical midline of the face).

FWHR Measurement and Data Analyses

The standard fWHR [(facial width)/(height from eyebrow to upper lip)] was tested for

significant differences between the experimental conditions by means of an ANOVA with

repeated measures. For all three methods alike, a picture was eliminated from the analyses

when both raters had independently coded the same type of problem with fWHR assessment

(e.g., both coded misaligned eyebrows or misplaced cheekbones). This affected a total of 30

pictures from 28 participants (listed in supplementary materials S2). Because experimental

condition (2 levels) as well as picture generation method (3 levels) amounted to six within-

subjects measures, a relatively high number of 21% data sets were incomplete (i.e., they were

missing at least one measure), out of a total of $N = 133$, slightly more than we had

preregistered as the maximal dropout rate for running a global ANOVA. Thus, we proceeded,

as preregistered, by analyzing data from each generation method separately, on samples with

slightly varying sizes but higher numbers of cases. Specifically, we ran three separate

ANOVAs, one per method, with condition as the only within-subjects factor. Bonferroni

corrected alpha error levels of $p = .017$ (.05 divided by three; note the typo in the

preregistration) were used for tests of significance. We chose ANOVAs instead of paired t-

tests because this provided us with the opportunity to compare the results with ANOVAs with

covariates (see Study 1b). Numbers (%) of participants that had to be excluded for these

analyses were 6 (5%) in the assembly-kit, 10 in the photo-edit (8%), and 13 (10%) in the

drawing task. The excluded images are listed individually for the aggressive-dominant and

peaceloving-submissive conditions in S2 of the supplementary materials. We note, however,

that results of the study would not have changed in any relevant way had we used more strict

or less strict exclusion criteria (supplementary materials S3).

All materials, scripts, and data can be found at https://doi.org/10.17605/osf.io/mtzax.
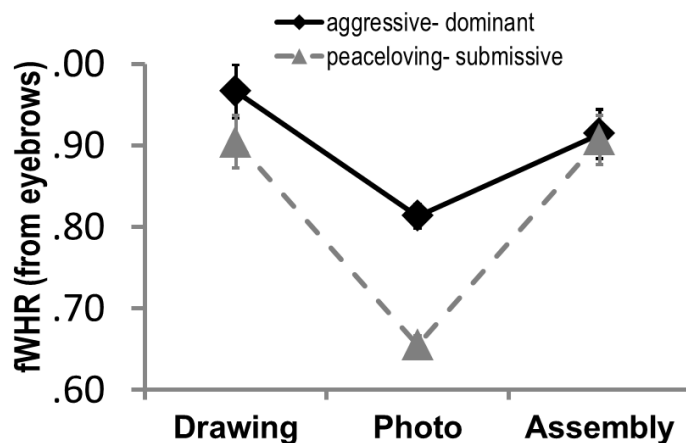
Results

The fWHR measures of the two raters showed high correspondence as indicated by

Pearson correlations of .923, .969 and .988 in the aggressive-dominant condition, and .870,

.945, and .999 in the peaceloving-submissive condition, for the drawing, photo-edit, and

assemply-kit methods, respectively.

Key results

Results of the ANOVAs comparing the aggressive-dominant with the peaceloving-

submissive faces showed no significant effect of condition for the drawings, $F(1,119) = 3.74$,

$p = .056$, $\eta^2 = .030$, or for the assembly task, $F(1,127) = 0.98$, $p = .754$, $\eta^2 = .001$, but a large

and highly significant effect of condition for the photo-edited images; $F(1,122) = 54.70$, $p <$

$.001$, $\eta^2 = .31$, as faces in the aggressive-dominant condition had higher fWHRs; $M = 1.80$,

95% CI [1.77, 1.83] than did faces in the peaceloving-submissive condition, $M = 1.66$, 95%

CI [1.64, 1.68]. Figure 1 shows results for all three methods, and sample pictures can be seen

in the Appendix. All images are available at https://doi.org/10.17605/osf.io/mtzax.

Figure 1

Mean fWHR of the faces in the two experimental conditions (aggressive-dominant and

peaceloving-submissive) generated in the drawing, photo-edit, and assembly tasks.

Note. Error bars represent standard errors of the mean.

Exploratory analyses

Consistency of the three items assessing individual prejudicing was relatively low due to the low number of items; Cronbachs α = .58. We correlated each score individually with the fWHR effect, i.e., the difference in fWHR between the aggressive-dominant and the peaceloving-submissive conditions. All correlations were insignificant and below r=.12. There was also no significant association of fWHR differences between conditions and the satisfaction, naturalness, and attractiveness ratings, averaged across conditions, as indicated by three separate multiple regression analysis, one for each method. All coefficients were insignificant with $p > .139$. Means and 95% CI of the ratings for all conditions and methods are given in the supplementary materials S4. Further exploratory analyses are provided in S5.

Discussion

We had hypothesized that if there is a stereotypical bias to associate aggressive-dominant personality traits in men with wider faces, participants would model faces with higher fWHR in the aggressive-dominant condition compared to the peaceloving-submissive condition in all three tasks. We did indeed find the expected result, but only for the photo-edit

method. The drawings showed only a weak and non-significant difference, and there was practically a null effect in the images from the assembly-kit method.

At least two factors might have contributed the results pattern. First, it might take a sufficiently realistic facial input to activate stereotypical beliefs about the association of character trait with facial shape. This type of input was provided in the photo-edit method, but not in the other two tasks. Thus, the difference between methods may be related to the higher involvement of bottom-up driven processing in the photo-edit method compared to the other two tasks. In the drawing method, the faces were only gradually developed, i.e., not presented up front, and their naturalness/"realisticness" depended on the drawings skills of the participant, which resulted in much higher fWHR variance compared to the other two tasks (Figure 1). In the images from the assembly task, the faces appeared more symbolic than realistic, and somewhat feminine due to the sharp contours of the provided shapes (as if accentuated by make-up), despite the light-grey shading that we had used to lessen this impression (see Appendix). Anecdotally, we note that these strong contours may also have caused participants to choose feature elements that were too small for the selected basic facial shape so that a relatively large number of the configured faces ended up looking quite unnatural.

Second, and more importantly, we observed that a number of the modeled faces quite clearly expressed emotional states, usually angry in the aggressive-dominant condition (V-shaped eyebrow arrangements, stiff lips), and happiness/friendliness in the peaceloving-submissive condition (/\-shaped eyebrow arrangements, lips smiling). This was true particularly for the drawings, but also for the photo-edited faces, despite our instructions in both tasks to model only neutral faces. Only the assembly-kit faces were free from differences between conditions in emotional state expression, since this method did not allow participants to tilt, reshape, resize, or otherwise change their form. To shed light on this issue,

we conducted another experiment (Study 1b) investigating the intensity and impact of

modeled emotional state expressions on the experimental effects found in Study 1a.

Study 1b

A sample of Bachelor students of Psychology at the end of their 3rd semester, when they had

already learned about the six basic emotions according to Ekman and Friesen (2003), and to

distinguish between emotional states and personality traits, evaluated whether the faces

generated by the photo-edit and the drawing methods in Study 1a showed a person in a

neutral state ("as in a biometric passport") or in some emotional state. We then counted the

number of emotionality judgments obtained for each image and entered this number into a re-

analysis of the data from Study 1a. In addition, we determined whether fWHR with height

taken from the nasion instead of the eyebrows would yield different results than those

reported in Study 1a. Because the position of the nasion is quite independent of facial

muscular activation, this measure should be less prone to emotional state expression (Kramer,

2017). The study is preregistered at https://doi.org/10.17605/osf.io/j682g.

Methods

Participants

Thirty-one students of Psychology (28 undergraduate), 22 female, 1 diverse, were

recruited via flyers on campus and via social media groups (one more than preregistered).

None of them had participated in Study 1a. We chose a full-sized sample of raters rather than

just a few, as customarily done, first, to obtain a more reliable estimate of state-emotionality,

and second, because we took a two-alternative forced choice instead of the usual Likert-type

rating response (but see Studies 2 and 3 where the opposite was done). Participants received

either course credit or € 8 for participation (ca. 45 minutes). Power considerations relate to

the sample of Study 1a because the inferential statistics will again be run on the images

generated therein. According to GPower (Faul et al., 2007), a unifactorial ANOVA with two

uncorrelated dependent measures will require $N = 106$ to detect a medium effect sized $f = .25$ with a power of 95% at an alpha level of .05. For correlations, $N = 115$ is required to detect a medium-sized correlation of .30 at $\alpha = .05$ and $\beta = 05$. Actual numbers were around 120 depending on method (see Study 1a); unsuitable images (because of lacking eyebrows, no frontal orientation etc) were eliminated in advance of the assessment.

Materials and Procedures

Participants were seated in front of a 22" computer screen in our lab. Up to three participants were investigated at the same time. They gave written informed consent. They were shown the drawings and the photos generated in Study 1a on full size display in randomized order, one by one, starting with the photos. Of the 544 pictures, 18 drawings and 3 photos were not shown because it was clear from Study 1a that these faces needed to be excluded, e.g., because they lacked eyebrows, appeared non-human, or were not frontally oriented.

Participants were told that the faces they would be shown belonged to different personalities, who were supposed to show a neutral state expression, as in a biometric passport photograph, but who sometimes nonetheless expressed an emotional state. They were asked to spontaneously press "p" on the keyboard for faces with a neutral expression, or "e" for faces that expressed an emotional state. In the latter case, participants were then transferred to another page where they were asked to code the specific type of emotion they had identified by entering a one-digit code number, as instructed on the screen. Reaction times were taken but not analyzed.

After completing the task, participants were asked four standardized questions about how they had experienced the study setting (for control purposes), and enter some demographical data. Finally, they were thanked and debriefed.

Data analysis

For each image, we counted how many raters had judged a face to display an emotional state as opposed to an emotionally neutral state, yielding numbers between 0 and 31. Numbers were then averaged across the two conditions for inferential analyses involving both experimental conditions.

As an alternative measure, we counted only state emotion judgments that conformed qualitatively to the character condition in which the image had been created, namely anger in the case of aggressive-dominant and happiness/fear/sadness in the case of peaceloving-submissive. We call these "trait conform" emotional states. Naturally, this number was lower than the count of emotions in general. For the drawings (140 images x 2 conditions), numbers were 90 vs. 42 in the aggressive-dominant condition (meaning, 42 of the 90 images judged state-emotional were "angry"), and 88 vs. 81 in the peaceloving-submissive condition. For the photos (123 x 2 conditions), the numbers were 95 vs. 82 in the aggressive-dominant condition, and 87 vs. 80 in the peaceloving-submissive condition. More details are provided in the supplementary materials S6.

In the inferential analyses reported here, we used the all-encompassing state-emotionality measure, as this is the less presumptional choice which also provides a larger range of values. Note, however, that we found practically the same results pattern with the more restrictive "trait-conform" measure (see S7 and S8 of the supplementary materials), which was less normally distributed (see especially S7A).

We performed the analyses numbered 1., 2., and 3. in the preregistration with the following two changes: First, to avoid any loss of data resolution, we used the full continuity of the state-emotionality measure in correlational analyses instead of creating a categorical variable with bins of low, medium, and high state-emotionality. Second, we did not have enough non-emotional faces to run the ANOVA preregistered under 1.; we had only 9 pairs of drawings and 12 pairs of photo-edited images with zero state-emotionality judgments (a

minimum of 30 image pairs had been preregistered). Thus, we ran an ANCOVA with

repeated measures and state-emotionality (averaged across the two conditions) as covariate,

and compared the results to those of the ANOVA in Study 1a.

Note that our use of repeated measures ANCOVA is conceptually broader than an

ANCOVA used for adjusting either between-group differences in randomized trials or

individual (pretest) differences in pre-post comparisons (Delaney & Maxwell, 1981). Our

design was meant to compare fWHR in the two within-subjects conditions while controlling

for between-subjects' differences in the application of state-emotionality (i.e., averaged

across both conditions). We were interested in the main effects of state-emotionality on

individual differences in mean fWHR *as well as* in the interaction of state-emotionality with

the fWHR difference between conditions. Thus, when we found significant effects of the

covariate, we plotted the regression of the within-subjects effect (fWHR differences between

conditions) on the covariate (state-emotionality averaged across conditions, varying from 0 to

31). In addition, we examined condition-specific correlations of the covariate and fWHR to

clarify the origin of the effects, thereby identifying potentially asymmetric (i.e.,

inhomogeneous) correlations across levels.

As preregistered, all analyses were re-run with height in fWHR taken from nasion

instead of eyebrows. All materials, scripts, and data can be found at

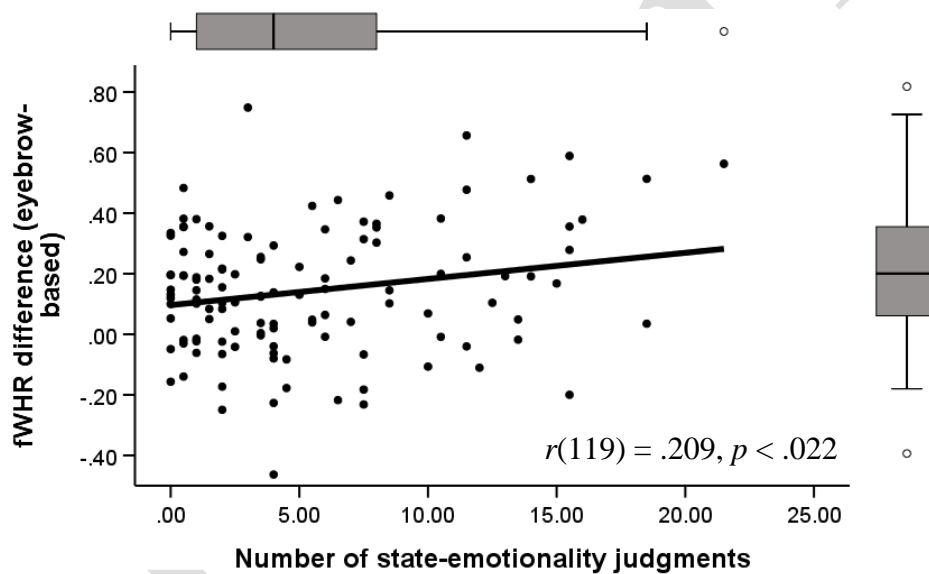https://doi.org/10.17605/osf.io/mtzax.

Results

Photo-edit method

Analysis 1. The ANCOVA of fWHR with condition as repeated factor and state-

emotionality as covariate showed a significant difference between the aggressive-dominant,

$M = 1.80$, 95% CI [1.78, 1.84], and the peaceloving-submissive condition, $M = 1.66$, 95% CI

[1.63, 1.68]; $F(1,118) = 12.49$, $p < .001$, $\eta_p^2 = .096$. However, compared to Study 1a, the

effect size (formerly $\eta^2 = .31$) was reduced to about one-third. There was also a significant

effect of state-emotionality, $F(1,118) = 7.77$, $p = .006$, $\eta_p^2 = .062$, and a significant interaction

of condition by state-emotionality, $F(1,122) = 5.38$, $p = .022$, $\eta_p^2 = .044$. Figure 2 shows the

associations.

Figure 2

Correlation of state-emotionality (as judged by the raters in Study 1b) in the photo-edited

faces with the effects of character on fWHR (i.e., difference between conditions in Study 1a),

with height taken from the eyebrows

Note. Shown on the abscissa are the number of raters that judged the images created by one

particular participant of Study 1a to convey state-emotionality (averaged across the two

conditions). In total, 182 faces (out of 246) were classified state-emotional. Error bars

represent 95% CI.

Analysis 2: The effect of condition on fWHR obtained in Study 1a (i.e., the difference

in fWHR between the aggressive-dominant and the peaceloving-submissive condition) was

significantly associated with the state-emotionality of the facial expressions (averaged across

conditions), but only when height was taken from the eyebrows, not from the nasion; this can

be seen by comparing the regression plots depicted in Figures 2 and 3.

Figure 3

Correlation of state-emotionality (as judged by the raters in Study 1b) in the photo-edited

faces with the effects of character on fWHR (i.e., difference between conditions in Study 1a),

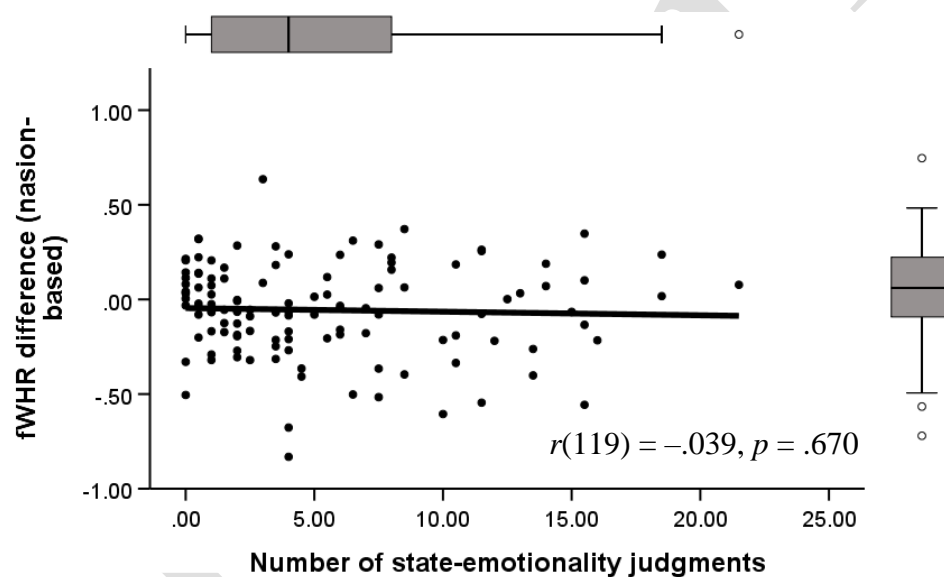*with height taken from the nasion*

Note. Shown on the abscissa are the number of raters that judged the images created by one

particular participant of Study 1a to convey state-emotionality information, averaged across

the two conditions. Details on the numbers are provided in the supplementary material S6.

Error bars represent 95% CI.

Analysis 3: As detailed in the first row of Table 1, fWHR correlated significantly with state-emotionality in the aggressive-dominant condition, but not in the peaceful-submissive condition. No significant correlations were found when height was taken from the nasion (rightmost coloumn of Table 1).

Table 1

Pearson correlation of fWHR of the photo-edited faces with state-emotionality judged by raters

Correlated variables	fWHR with height from eyebrows	fWHR with height from nasion
fWHR, state-emotionality	$r_{agg}(118) = .273, p = .003$	$r_{agg}(118) = .016, p = .859$
	$r_{peace}(118) = -.035, p = .705$	$r_{peace}(118) = .078, p = .399$

Note. r_{agg} = Pearson correlation in the aggressive-dominant condition, r_{peace} = Pearson correlation in the peaceloving-submissive condition. Listwise elimination of missing data.

Drawings

For the drawings, the marginally significant effect found in the ANOVA of Study 1a turned clearly nonsignificant in the ANCOVA with state-emotionality as covariate, $F(1,118) = 0.49, p = .484, \eta_p^2 = .004$. The covariate itself was significant, $F(1,118) = 4.60, p = .034, \eta_p^2 = .038$. The effect went into the same direction as in the photos, but was by far less strong. The interaction of condition by state-emotionality was not significant, $F(1,122) = 1.54, p = .218, \eta_p^2 = .013$. Details are shown in Figures S7B, S7C (for eyebrows) and S8B, S8C (for nasion) and in Table S9 of the supplementary materials.

Exploratory Analyses

For both the drawings and the photos, fWHR with height taken from nasion was not significantly higher in the aggressive-dominant condition compared to the peaceloving-submissive condition, and showed no covariation with state-emotionality (supplementary materials S10).

Discussion

Results show that, according to the state-emotionality judgments obtained from the independent group of raters, the fWHR effects of Study 1a are indeed significantly confounded with emotional state information. For the photos, although the effects of condition on fWHR were still significant when state-emotionality was statistically controlled for, they were reduced by two-thirds in size. Our post-hoc correlational analyses showed that it is particularly the aggressive-dominant condition in which fWHR correlates significantly with state-emotionality, and that this relationship disappears when the nasion is taken instead of the eyebrows to determine the height of the face. Effects for the drawings went into the same direction, but were not significant.

It is interesting that participants reverted so strongly to modeling emotional states despite having been instructed to model only faces with neutral expressions. Due to the "phenotypic overlap" of emotional states with fWHR (Deska et al., 2018a), applying state-emotionality to the facial images increased the fWHR differences between the two trait conditions. However, this was true only for fWHR with height taken from the eyebrows, particularly in the aggressive-dominant condition, where anger was often expressed through frowned eyebrows, thereby reducing facial height and increasing fWHR (Costa et al., 2017; Deska et al., 2018a; Marsh et al., 2014). Justin Carré himself demonstrates this effect at https://www.innovation.ca/story/faces-aggression (latest access September 30th, 2020). This implies, however, that fWHR cannot be unambiguously interpreted in terms of bone structure

(testosterone-related or not) but is influenced by state-related, transient activation of facial

musculature (Hehman et al., 2015; Marsh et al., 2014).

Instead of using eyebrows, facial height is sometimes taken from the eyelids (Lefevre

et al., 2014; Stirrat & Perrett, 2010; Yang et al., 2018), but according to our results, the

nasion may be an even better choice (Kramer, 2017; Weston et al., 2007) because its vertical

position should be quite independent from muscular activations around the eyes, including

their clenching in connection with anger. In fact, when using the nasion instead of the

eyebrows to determine facial height, we did not find any significant influence of state-

emotionality on fWHR, not even in the photo-edited faces. However, this does not mean that

the remaining mean fWHR difference between the conditions is entirely free of state-

emotionality. It may as well mean that raters' judgments in Study 1b were not sensitive

enough to assess state-emotionality exhaustively, perhaps owing to the relatively gross two-

alternative (yes/no) categorization task that was used instead of a Likert-type rating task.

Alternatively, perhaps it is impossible in principle to distinguish between trait aggression and

state anger from 2D photographs.

A third possibility is that FaceGen© imposed certain technical limits so that

participants were misled in expressing their genuine ideas about the faces. For instance, they

may have used the sliders on top of the list more than those at the bottom (the reader may try

a comparable task with 37 sliders at http://facemaker.uvrg.org/#, latest access September 30,

2020). On the other hand, participants were entirely free in their feature choice in the

drawings where we found no significant effect of condition either. Yet in that method,

participants may have been limited by their drawings skills. Yet again, the assembly-kit

method, which was perceptually and response-wise the least difficult of the three tasks

(Sayõl, 2001; Verstijnen et al., 1998), also produced practically null differences between

conditions.

To zero in on the mediating roles of the drawing skills and the portrayal of state-emotionality in the two conditions, we conducted two further experiments with free drawings (Study 2) and the photo-edit method (Study 3) in which we asked participants more explicitly to refuse from portraying state-emotionality.

Study 2

We re-ran the experiment with the drawings from Study 1a for which the effect of condition had been nonsignificant (at $p = .056$). This time, we explained to participants the difference between character traits and state emotions, and asked them several times in the instructions to draw two characters as traits and with a "pokerface", a face that would mask any state emotions (c.f. study 1a by Deska et al., 2018a). We also asked participants to draw the faces as realistically as possible. The study is preregistered at https://doi.org/10.17605/osf.io/xma4t.

We had two naïve raters evaluate for each face on a scale of 0 to 6, first, how "natural" it appeared to them, and secondly, how "state-emotional" it appeared to them. All analyses were rerun for fWHR taken from the nasion.

Methods

Participants

We recruited 158 participants on campus, via social media, and/or among friends and acquaintances, 18 more than preregistered because a larger than expected number of students recruited in an introductory Psychology lecture signed up for participation, and we did not want to turn anyone down. None of the participants had participated in Studies 1a or 1b.

Seven of the participants indicated that they had heard of research linking facial shape with character traits before. Two more indicated that they had drawn a person that exists in real life in at least one of the conditions. Data of these participants were eliminated as preregistered. Data of 15 more participants were excluded in line with the preregistration

because the two raters had indicated identical problems with the images. Two data sets were lost due to experimenter error during fWHR rating.

The final sample consisted of 132 participants, mean age 28.47 years (82 female, 46 male, one diverse, three missing values for gender).

Materials and Procedures

Participants received a sheet of paper with study information and instructions. On the top of an attached sheet in landscape format, they provided informed consent, demographic data, and whether they knew of any research linking facial features to character. The area below was segregated by a vertical line to provide the spaces for the drawings of the face of an aggressive-dominant male and a peaceloving-submissive male "the way common sense would conceive of it" (with left and right position of condition counterbalanced between participants). Above the two drawing spaces, the reminders "ATTENTION: frontal, realistic, pokerface!" were written in boldfaced font and underlined. At the bottom of the sheet, participants were asked to indicate whether they had drawn the face of someone who exists in real life.

Participants who were recruited personally were also asked to indicate the position of the cheekbones of the faces after the drawings were completed. However, these marks were later ignored because they were given for only a subset of the faces and were placed at unrealistic positions in many cases.

FWHR measurement

Two raters (J. B. and O. F.) who had not participated at Studies 1a or 1b, one of them blind to the hypotheses and conditions of the study and the other one entirely naïve, measured facial width, facial height from nasion, and facial height from eyebrow (taking the approximate mean of the two if these were not aligned). Thereafter, the raters went through the pictures again one by one to indicate on a Likert scale of 0 (not at all) to 6 (very strong)

how natural/realistic a face appeared to them, and in a separate, second trial, using the same

Likert scale, how much state-emotionality they saw in a face. For the latter, they were

instructed to indicate rather spontaneously to what degree the face showed any emotional

state expression, regardless of what quality. They were further explained that state emotions

were current feelings expressed through muscular activation, as opposed to static facial

features that depended on bones, body fat, or identity of the person.

Interrater correlation of the state-emotionality ratings was $r(130) = .666$, $p < .001$, in

the aggressive-dominant, and $r(130) = .577$, $p < .001$, in the peaceloving-submissive condition.

For naturalness, the correlations were $r(129) = .821$, $p < .001$, and $r(130) = .827$, $p < .001$,

respectively. The two rating measures, averaged across conditions, correlated negatively,

$r(128) = -.423$, $p < .001$. This indicates that the higher the state-emotionality, the less natural

the faces appeared to the raters.

Data analyses

Inferential analyses were performed as preregistered; Bayesian analyes were added. All

materials, scripts, and data can be found at https://doi.org/10.17605/osf.io/mtzax.

Results

The ANOVA of the ordinary eyebrow-based fWHR showed no significant main effect

of the repeated measures factor condition, $F(1, 124) = 0.31$, $p < .578$, $\eta^2 = .002$. This was also

true for the preregistered ANOVA with emotional (state-emotionality > 0) and unnatural

(naturalness < 3) faces excluded, $F(1, 22) = 0.00$, $p < .988$, $\eta^2 < .001$. The pattern did not

change when instead of excluding images, state-emotionality and naturalness were used as

covariates in an ANCOVA; none of these analyses showed any significant effects, and are

therefore not detailed further. We did, however, find a significant effect of condition on

fWHR when height was taken from nasion, $F(1, 129) = 4.84$, $p < .030$, $\eta^2 = .036$. However,

the effect went into the *opposite* from the expected direction because fWHR was lower in the

aggressive-dominant condition, $M = 2.28$, 95% CI [2.20, 2.36], compared to the peaceloving-submissive condition, $M = 2.36$, 95% CI [2.28, 2.44].

Bayesian Analyses

Using JASP (2020, version 0.13.1, jasp-stats.org, Wagenmakers et al., 2018), we performed Bayesian analyses of the eyebrow-based fWHR data to quantify evidence for the null hypothesis compared to the alternative hypothesis. Compared to the null model that knows only the intercept, our data were estimated to be $BF_{01} = 6.24$ times more likely to occur under the null hypothesis than under the alternative hypothesis stating a repeated measures effect. A factor higher 3 has originally been interpreted as "strong evidence" (Jeffreys, 1961), but was later balanced out to "moderate evidence" (Lee & Wagenmakers, 2013). Note that this factor was obtained despite the fact that state-emotionality was not partialled out in this analysis.

Discussion

Participants in the study were free to draw the faces any way they wished based entirely on their imagination; they were not limited by particular input or output requirements. Yet they did not show any significant evidence of the idea that an aggressive-dominant man might have a higher fWHR compared to a peaceloving-submissive man. In fact, the Bayes factor suggests that the data are 6 times more likely to occur under the null hypothesis compared to the alternative hypothesis.

The effect of condition remained absent when drawing skills were entered into the equation by various means. When only the more realistic faces were analyzed (i.e., those evaluated "reasonably natural" by the two raters, above the mid-point of the scale), the fWHR difference between conditions remained close to zero. Entering the naturalness ratings at their full spectrum as a covariate into the ANCOVA of all data sets made no noteworthy difference either.

The analysis of the nasion-based fWHR did not show any relevant difference between conditions as the effects went into the opposite than expected direction, with wider faces observed in the peaceful-submissive condition compared to aggressive-dominant.

In Study 3, we proceeded to investigate the role of state-emotionality in the photo-edit method.

Study 3

We set out to conceptually replicate the findings of Study 1a obtained with the photo-edit method under somewhat varied experimental conditions. We used the same statistical prototype of a white European male aged 30 as template. However, instead of asking participants to model particular character traits, we now asked them to model faces of male individuals with high versus low *social status.* This focusses on social rank, a variable that has consistently been found to be related to fWHR (Goetz et al., 2013; Hahn et al., 2017; He et al., 2019; Noser et al., 2018), rather than on latent motivational traits. We additionally asked participants to model the faces with a "pokerface", as in Study 2, to minimize the modeling of emotional state expressions as much as possible. Contrary to Studies 1a, 1b, and 2, the sample contained no students of Psychology. The study is preregistered at https://doi.org/10.17605/osf.io/xma4t.

Methods

Participants

153 individuals who had not taken part at the other three studies were recruited to participate for € 10 and a small gift (a Goethe University pen and college block). Data of 11 participants were not processed further (i.e., their fWHR was not determined) because they indicated at the end of the session that they were students of Psychology. Five participants were excluded because participants indicated that they had modeled the face of a person who exists in real life in at least one of the conditions; two of these also indicated upon

questioning that they had heard before of research linking personality traits to facial shapes. Two data sets were lost during fWHR measurement due to experimenter error. Nine data sets were excluded because both raters had noted that they found the face to be extremely unnatural ("freaky").

The final sample consisted of 126 participants, mean age 27.25 years, 85 female, 46 male, 2 diverse, and 1 missing entry.

Procedures

Procedures were equivalent to those in Study 1a, except that instead of a second screen, participants had a large digital clock showing the remaining minutes (out of 10) and a cardboard reminding them of the current condition ("powerful" vs. "powerless") placed on the desk right in front of them. After creating the facial images, they filled out a computerized questionnaire regarding their individual tendency to prejudice (not considered further here). In another task, they were shown the photo-edited images from Study 1a one by one on the screen to re-classify whether it showed an aggressive-dominant or a peaceloving-submissive face (not considered further here). Finally, they indicated demographical data and whether they knew of any research linking facial shape to personality traits.

The same raters as in Study 2 measured fWHR, and thereafter rated naturalness and state-emotionality, as in Study 2.

fWHR Assessment

Intercorrelations between raters were $r(124) = .926$, $p < .001$, $r(124) = .977$, $p < .001$, for fWHR with height taken from eyebrows, and $r(124) = .873$, $p < .001$, $r(124) = .966$ $p < .001$, for fWHR with height taken from nasion. Interrater correlation of the ratings in the powerful condition was $r(124) = .611$, $p < .001$, for state-emotionality, and $r(124) = .604$, $p < .001$, for naturalness. In the powerless condition, these correlations were, $r(124) = .579$, $p < .001$, and $r(124) = .635$, $p < .001$, respectively.

With a mean rating of $M = 0.57$, 95% CI [0.40, 0.73] in the powerful condition and $M = .98$, 95% CI [0.79, 1.17] in the powerless condition (on a scale from 0 to 6), the ratings were quite low for state-emotionality, as expected, and were in fact higher in the powerless condition compared to powerful. Naturalness ratings were quite high, $M = 4.56$, 95% CI [4.33, 4.80] for powerful, and $M = 3.51$, 95% CI [3.23, 3.80] for powerless. As in Study 2, state-emotionality and naturalness correlated highly negatively, $r(124) = -.637$, $p < .001$, in the powerful condition, and $r(124) = -.510$, $p < .001$, in the powerless condition.

Data analyses

Data analyses were identical to those described in Study 2. Exploratory analyses, Bayesian analyses and an equivalence test were added. All materials, scripts, and data can be found at https://doi.org/10.17605/osf.io/mtzax.

Results

The ANOVA of fWHR showed a significant main effect of the repeated measures factor condition, $F(1,126) = 27.59$, $p < .001$, $\eta_p^2 = .18$. Mean fWHR was higher in the powerful condition, $M = 1.93$, 95% CI [1.91, 1.95], compared to the powerless condition, $M = 1.83$, 95% CI [1.80, 1.86].

To exclude emotional and unnatural faces, as preregistered, we used only images with a state-emotionality rating of 0 and a naturalness rating higher than 2. The 29 image pairs that fulfilled these criteria showed no significant effect of condition $F(1, 28) = 0.94$, $p = .343$, $\eta_p^2 = .03$. Mean values were $M = 1.93$, 95% CI [1.90, 1.97] in the powerful condition compared to $M = 1.90$, 95% CI [1.85, 1.95] in the powerless condition.

As in Study 1b, an ANCOVA was run with state-emotionality and naturalness as covariates to inform about the continuous effect of these variables on condition-dependent fWHR variation. In this analysis, the previously significant effect of condition turned insignificant, $F(1,123) = 1.95$, $p = .166$, $\eta_p^2 = .016$, while the effect of state-emotionality was

highly significant $F(1,123) = 12.03$, $p = .001$, $\eta_p^2 = .09$. Naturalness had no significant effect, $F(1,123) = 0.62$, $p = .433$, $\eta_p^2 = .005$. Interactions were not significant (at $p = .05$).

As preregistered, we also analyzed fWHR with height taken from the nasion. This ANOVA showed a just-significant effect of condition, $F(1,125) = 4.309$, $p = .040$, $\eta_p^2 = .033$, as fWHR was significantly higher in the powerful condition, $M = 2.15$, 95% CI [2.13, 2.18], compared to the powerless condition, $M = 2.11$, 95% CI [2.07, 2.14]. The effect turned nonsignificant in an ANCOVA with state-emotionality and naturalness as covariates, $F(1,125) = 3.44$, $p = .066$, $\eta_p^2 = .027$, with no significant interactions (at $p = .05$).

Exploratory analyses

To explore the effects of state-emotionality specifically on the positioning and orientation of the eyebrows, we computed fWHR with height taken from the distance between the (malleable) brow and the (static) nasion (not preregistered). The repeated measures ANOVA showed a significant fWHR difference between the powerful, $M = 25.17$, 95% CI [20.93, 29.42], and the powerless condition, $M = 14.34$, 95% CI [10.00, 18.68]; $F(1,125) = 12.01$, $p = .001$, $\eta^2 = .09$. In the ANCOVA with state-emotionality as covariate, the effect of condition was not significant; $F(1,124) = 0.49$, $p = .484$, $\eta_p^2 < .01$; nor was there a significant main effect of state-emotionality; $F(1,124) = 2.95$, $p = .088$, $\eta_p^2 < .02$. However, the interaction effect was highly significant; $F(1,124) = 21.53$, $p < .001$, $\eta_p^2 = .15$. The regression plot is shown in the supplementary materials S11. The correlation of the fWHR difference between conditions with state-emotionality averaged across conditions was significant in the powerful condition, $r(124) = .387$, $p < .001$, and a trend was observed in the powerless condition, $r(124) = -.170$, $p = .057$. Thus, state-emotionality appears to have increased fWHR by lowering the vertical positioning of the mid-eyebrows relative to the nasion in the powerful condition, and tended towards the opposite in the powerless condition where fear and sadness were sometimes expressed by /\-type of eyebrow arrangements.

Bayesian analyses

For each of the fWHR measures, we estimated Bayes factors to estimate how likely the differences between the two repeated conditions are to occur under a null hypotheses as compared to the alternative hypothesis when state-emotionality was considered as covariate. Results of the Bayesian repeated measures ANCOVA performed by JASP (2020, version 0.13.1, jasp-stats.org, Wagenmakers et al., 2018), are detailed in the supplementary materials S12. We left naturalness out of these analyses because the variable had not yielded any significant effects anyway, and was highly anticorrelated with state-emotionality anyway.

For the analysis of the eyebrow-based fWHR measure, we included the covariate in the null model to partial out the effects of state-emotionality. Results showed that compared to the model containing only the repeated measures factor Condition, the null model was $BF_{01} = 672,790.15$ times more likely to occur under H0 compared to H1. This is extremely strong evidence for H0 compared to H1 with the main effect of the covariate partialled out. The corresponding Bayes factor for the nasion-based fWHR measure was only $BF_{01} = 1.59$, suggesting a comparably low impact of state-emotionality in driving the effects of condition towards null.

For our exploratory fWHR measure that uses the distance from eyebrows to nasion as a measure of height, we performed a Bayesian linear *regression* analysis to be able to capture the interaction effect that was significant in the classical ANCOVA reported above. To this end, we regressed the fWHR difference between conditions on the covariate state-emotionality. Results estimated the null model (i.e., intercept with no factor specified) to be $BF_{01} = 1943$ times more likely to occur under H0 compared to H1 than the model that contains state-emotionality as covariate. Again, this shows how strongly the fWHR differences between the two trait conditions depended on participants' application of state-emotionality to the faces.

Equivalence test

We investigated the practical equivalence of our findings to test for absent effects of condition when state-emotionality was controlled for. We used only facial width in this analysis, without correcting for height, to exclude any expression-related information that is based on soft-tissued features like eyebrows or lips.

The equivalence test requires specification of the SESOI, the *smallest effect size of interest* (Harms & Lakens, 2018). In our case, this would be the smallest difference in facial width that signals aggressive-dominant as compared to peaceloving-submissive traits in the absence of any information on state-emotionality.

Using data from an online experiment (data available at https://doi.org/10.17605/osf.io/mtzax, under *Study 3/Equivalence Test*), we estimated the SESOI based on responses of $N = 110$ independent observers who were presented with a subset of the photo-edited images obtained in Study 1a, and asked to re-categorize each face into either the aggressive-dominant or the peaceful-submissive condition. All images were selected for the study that showed an (eyebrow-based) fWHR difference between conditions exceeding the median fWHR difference of .105, and a state-emotionality rating of <=0.5 (as obtained in Study 1b). In addition, facial expressions were fully masked on all images (using the identical mask for all images), so that the reclassification had to be performed solely based on facial shape information. We computed reclassification accuracy as hit rate minus false alarm rate, termed P_r according to two-high threshold theory of recognition (Snodgrass & Corwin, 1988).

We determined the difference in facial width (in pixels) at which P_r significantly differed from 0, analogue to a "just noticeable difference" or "above-chance discrimination" threshold. After eliminating one case with extremely negative P_r (-.27), and two cases with missing fWHR, the remaining 11 pairs of images were accurately reclassified at P_r = .101,

95%CI [-.001, .203], just significantly above zero; $t(10) = 2.20$, $p = .052$), rendering

permutation tests obsolete. Although 11 image pairs is a relatively small set, our stimuli were

highly standardized because all faces originated from the same template.

The difference in facial width associated with the just significant P_r was $M = 8.55$,

95% CI [2.21, 14.87]. This difference reflects the SESOI; the smallest facial width difference

that allows for discrimination between trait conditions just above chance level when feature-

based height information is masked (and state-emotionality eliminated). This value is to be

compared with the upper bound of the 95% CI of the actually observed mean differences in

facial width.

Observed estimated mean differences in facial width, adjusted for the effects of state-

emotionality, were $M = 0.60$, 95%CI [-4.34, 5.54], $\eta_p^2 = .026$, which clearly falls below the

SESOI. This was true not only for the photo-edited faces in Study 3, but also for those of

Study 1a, where the estimated adjusted mean difference between conditions was even

negative, $M = -3.89$, 95%CI [-6.41, 1.36]. Note, however, that statistically adjusting for state-

emotionality made hardly any difference for the shape-based facial width measure, in line

with our presumptions.

We conclude that facial width had not been modeled to signal trait to a practically

meaningful degree in Studies 1a and Study 3; what gave rise to the significant trait

differences in fWHR in these two studies was only (expression-related) height information.

Discussion

At first blush it seemed that the photo-edit method in this study had produced the expected

fWHR differences between the conditions, contrary to the free drawings obtained in Studies

1a and 2. Faces of powerful men were modeled with higher fWHR compared to faces of

powerless men. However, the effect disappeared when faces later judged as expressing state-

emotionality were eliminated from the analyses, or when state-emotionality was statistically

controlled for. In fact, Bayes analyses and the equivalence test clearly speak for a null effect of condition on fWHR when state-emotionality is controlled for.

Exploratory analyses on a novel fWHR measure that takes facial height as the difference between eyebrows and nasion suggested that the effect of state-emotionality was mainly caused by lowered ("frowned") eyebrows in the powerful condition, and a tendency towards heightened ("raised") eyebrows in the powerless condition. This result is consistent with prior studies (Costa et al., 2017) showing that facial height more than width conveys emotion-related information via fWHR. On the other hand, it demonstrates that fWHR is no unequivocal size-independent measure for cheekbone distance (Hehman et al., 2015). Instead, fWHR varies with muscular activations, especially with those underlying eyebrow orientation during emotion expression (Ekman & Friesen, 2003; Tipples et al., 2002; Witkower & Tracy, 2020).

General Discussion

We investigated observer's ideas about the facial width-to-height ratio of aggressive-dominant (Studies 1a and 2) and powerful (Study 3) males in comparison to their character opposites, respectively. By having participants draw, assemble, or photo-edit faces of the specified characters, we investigated the degree to which these personality traits are linked with males' fWHR in observers' imagination. Within social perception-action-reaction cycles, such beliefs may trigger, maintain or strengthen aggressive tendencies in targets with relatively high fWHR, above and beyond biological predispositions, on account of interactive, spiraling dynamics (Jaeger et al., 2019; Loeb et al., 2016; Yildirim & Derksen, 2012; Zayas et al., 2002).

The methods that participants were given in the present studies to express their ideas of the pre-specified characters' faces varied at the input level as well as the output level. The given stimulus input varied from entirely lacking in the drawing task over disassembled,

sketchy shapes in the assembly-kit method to realistic facial images in the photo-edit method (c.f. Tipples et al., 2002). The output modality varied from free externalization (drawing) over selection-construction (assembly) to configurally constrained feature manipulation via digital sliders (photo-edit). We expected the relative proportion of top-down controlled processing relative to bottom-up driven activation to co-vary with these variations (Dijkstra et al., 2019, Fan et al., 2020; Michelon & Koenig, 2002). We figured that a more realistic stimulus input and a more constrained output would make it easier for participants with relatively poor perceptual imagery and drawing skills to activate and express their ideas.

We hypothesized that participants would generate wider faces in the aggressive-dominant and powerful conditions compared to their contrasting comparison conditions, respectively, in all three tasks. Contrary to the expectations, we found practically no significant differences between conditions in assembly-kit images and in the drawings, the two tasks that did not provide truly realistic facial stimulus input for participants to work with. fWHR differences between conditions were very small for both methods (and practically null for the assembly-kit) despite the test power of .95. Although there was some evidence of an association of fWHR with character condition in the expected direction in the drawings of Study 1a (at $p = .056$), further analyses showed that this association did not emerge from facial shape variation, but from displays of state-emotionality. Specifically, the evidence collapsed when the contribution of state-emotionality i. was statistically controlled for (Study 1b), ii. was discouraged by the pokerface instruction (Study 2), or iii. was minimized by using the nasion instead of the eyebrows to measure facial height (Kramer, 2017; Weston et al., 2007). In addition, Bayesian analyses of the data of Study 2 suggested that the observed fWHR differences are more than 6 times more likely to occur under the null hypothesis compared to the alternative hypothesis. Thus, we conclude that participants have

not modeled fWHR to signal aggressive-dominant as compared peaceloving-submissive

character traits in any of the two techniques, at least not to a noteworthy degree.

In the photo-edit technique where a realistic, near-photo quality template of a male's

face was to be modified, strong effects of character trait condition seemed to exist at first

blush. In Study 1a, faces in the aggressive-dominant condition were modeled with a much

higher fWHR than faces in the peaceloving-submissive condition. In Study 3, the same was

found for the powerful condition compared to the powerless condition. Effect sizes were very

large, with Cohen's d of about 1 and above. The high contrast was observed as the

comparison conditions were semantic opposites, with both deviating from "null" or "normal"

(c.f. Zhang et al., 2020). Notably, however, we found no significant effect of order of

condition, and no interaction of order by condition, in the two studies with the photo-edit

method (Study 1a and Study 3, see supplementary materials S13); in fact, these interaction

effects were near zero in size. This suggests that participants modeled the faces individually

to match the current condition instructions, and not to maximize the contrast between the two

conditions. That said, however, correlation patterns in Study 1a suggest that the aggressive-

dominant condition may have contributed more to the fWHR difference between conditions

than did the peaceloving-submissive condition; a pattern that was less clear in Study 3.

At this point, the strong fWHR differences between the conditions in the photo-edit

method seemed to confirm the existence of a social attribution bias that associates aggressive,

dominant, and powerful personality traits compared to their character opposites with wider-

shaped male faces. The fact that the effect emerged only under task conditions that involved a

photo-realistic facial input, contrary to the drawing and assembly-kit methods where less

detailed perceptual input has been given, would then have to be interpreted as evidence that

sufficiently detailed bottom-up driven processing is needed to activate the stereotypical

association.

A closer look, however, revealed that this interpretation would be erroneous because the effects in the photo-edit method were, and practically at full size, mediated by modeled state-emotionality. The evidence for this conclusion is fivefold: First, the effects of character condition became considerably smaller in Study 3 compared to Study 1a, when the added "pokerface" requirement discouraged the modeling of state-emotions. Second, the effects were much weaker (Study 3) or even clearly insignificant (Study 1b) when the nasion was used instead of the eyebrows to measure facial height. Third, the analyses of only non-emotional faces showed practically a null finding (Study 3), although, admittedly, on a relatively small subsample of faces ($n = 29$). Fourth, when the use of state-emotionality was statistically controlled for by ANCOVA, the effects of condition were either markedly reduced (Study 1b), or non-significant (Study 3). Finally, and perhaps most importantly, Bayesian analysis of the data from Study 3 estimated that, when state-emotionality was controlled for, the null hypothesis was thousands of times more likely to occur under the observed data pattern than the alternative hypothesis. The same conclusion was drawn from the equivalence tests performed for the data of Study 3 and of Study 1a which also showed that in the absence of any soft-tissue based expression of state-emotionality, the observed mean differences in facial width were practically equivalent to zero.

Exploratory analyses on the data of Study 3 suggested, essentially, that the positioning/orientating of the eyebrows relative to the nasion was the main source of the fWHR variation between conditions. In particular, fWHR was higher in the powerful condition where the image of an angry man was relatively often created via lowered/frowned eyebrows, thereby reducing the facial height measure and increasing fWHR (c.f. Tipples et al., 2002; Witkower & Tracy, 2020). A tendency towards the opposite was observed in the powerless condition, where fear and sadness were sometimes expressed through raised or /\-shaped eyebrow arrangements. Compared to the other two methods, this simple and

superficial approach to solving the character generation task may have been encouraged by the photo-edit method due to the prominent positioning of the eyebrow sliders in FaceGen© which lists eyebrow sliders on top, cheekbone sliders directly below, and other features further below. By modifying the position, shape, and orientation of the eyebrows, participants created facial expressions judged later by independent raters to convey information on state-emotionality, despite having been instructed (repeatedly) not to do so, and despite having been instructed to model only stable character traits or social status.

When personality trait information is confused with emotional state information, the literature speaks of emotion overgeneralization (Said et al., 2009; Todorov et al., 2015; Zebrowitz, 2017; Zebrowitz et al., 2010; Zebrowitz & Montepare, 2008). In the case of state anger, emotion overgeneralization means that individuals who happen to have, for instance, relatively low-positioned eyebrows by nature, may be more likely to be treated habitually by other people as if they were constantly ill-tempered compared to individuals with relatively high-positioned eyebrows (Deska et al., 2018a, 2018b; Hehman et al., 2015; Marsh et al., 2014; Said et al., 2009). To the degree that such treatment in turn triggers adverse reactions, hostility might spiral, and repeated such experiences may be internalized (Elshout et al., 2017; Jaeger et al., 2019; Loeb et al., 2016; Yildirim & Derksen, 2012; Zayas et al., 2002). In effect, trait aggressiveness inferred by the observer from low-positioned eyebrows may have the same noxious effects as biologically anchored antisocial traits on the side of the target (Marsh et al., 2014). What might begin as a motivationally irrelevant peculiarity of one's facial shape could give rise over time to a habitual tendency to respond aggressively to the perpetual encountering of skepticism and distrust by other people.

From a scientific point of view, it is important to tease apart the effects of (observer-driven) emotion overgeneralization from (target-driven) hormone- or otherwise biologically mediated associations between fWHR and character trait. In segregating the latter from the

former, one should better rely on bone-based indices of facial width and height as in skull measurements (Kramer, 2017). Alternatively, and confirmed by our results, the nasion may be a valid alternative to the (emotion-sensitive) measure of eyebrow position in assessing facial height. A third option would be to provide target-specific reference frames for facial expression dynamics prior to having observers infer trait information from static images; for example, they could watch the target person speak, sing, or display various state emotions before evaluating the trait implications of their facial morphometry. However, it may not be possible to fully tease apart state and trait information from static 2D images alone after all. Dynamic facial information allows for more accurate and valid inferences regarding both emotional states and personality traits (Krumhuber et al., 2007; LaBar et al., 2003; Trautmann et al., 2009; Weyers et al., 2006).

Building on the present interpretations and those of prior studies (Deska et al., 2018a; Marsh et al., 2014), the role of emotion generalization could be investigated further. First, existing data sets that have previously been interpreted in terms of the testosterone hypothesis may be reanalyzed with a bone-based height measure such as the nasion. This might reduce the previously reported fWHR effects, most of which are eyebrow-based (Haselhuhn et al., 2015). Second, any observed associations between estimated aggression/dominance/power and eyebrow-based fWHR should become smaller when emotion overgeneralization is being prevented as much as possible by familiarizing observers with individual facial dynamics of target individuals. Finally, our study design with the photo-edit method showing strong effects of emotion overgeneralization could be applied to facial templates of females (Martin et al., 2019). Since fWHR according to the testosterone hypothesis is presumed to be sexually dimorphic (although disputed, e.g., by Kordsmeyer et al., 2019; Kramer, 2017), while emotion overgeneralization should affect both sexes equally, this variation could help to dissociate the social and the biological factors. In fact, plastic surgery, being performed

mostly on women, warns against malalignment of eyebrows for its unfavorable impact on

social attributions (Yalcinkaya et al., 2016).

Conclusion

We conclude from our extensive analyses of the images generated with all three

methods that aggressive-dominant traits and high social status is linked with high fWHR in

people's minds only to the degree that state-emotionality is involved. State emotions were

expressed, in particular, by unnatural arrangements of eyebrows, whereby the impression of

anger was created for an aggressive-dominant man, while fear, sadness, or friendliness was

used for a peaceloving, socially inferior man. In the absence of state-emotionality, differences

in fWHR were close to zero in all three image types. We infer, first, that observers show no

practically relevant attribution bias linking aggressive-dominant personality traits to fWHR in

males, and secondly, that overgeneralization of expressed emotions should be considered

more when designing and interpreting studies linking fWHR with behavioral and

motivational traits.

Our conclusions are not directly at odds with the classical interpretation that wider-

faced men are indeed more aggressive and dominant than narrow-faced men, although we

wonder why this link was not represented in the created images. By contrast, our results are

perfectly consistent with the idea that such antisocial dispositions may be socially *induced* by

the discriminative treatment that men whose facial features happen to resemble expressed

anger experience across their lifetimes (Marsh et al., 2014). Evidence from lab (Matsumoto &

Hwang, 2021) and field studies (Wilson & Rule, 2015) suggesting that men with high fWHR

might get unrightfully condemned of crimes more easily than do those with low fWHR

demonstrate how plausible and realistic this idea is.

References

Abraham, A. (2018). *The Neuroscience of Creativity*. Cambridge University Press.

Ahmed, S., Sihvonen, J., & Vahamaa, S. (2019). CEO facial masculinity and bank risk-taking. *Personality and Individual Differences*, *138*, 133-139. https://doi.org/10.1016/j.paid.2018.09.029

Alrajih, S., & Ward, J. (2014). Increased facial width-to-height ratio and perceived dominance in the faces of the UK's leading business leaders. *Br J Psychol*, *105*(2), 153-161. https://doi.org/10.1111/bjop.12035

Anderl, C., Hahn, T., Schmidt, A. K., Moldenhauer, H., Notebaert, K., Clement, C. C., & Windmann, S. (2016). Facial width-to-height ratio predicts psychopathic traits in males. *Personality and Individual Differences*, *88*, 99-101. https://doi.org/10.1016/j.paid.2015.08.057

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269-278. https://doi.org/10.1037/1528-3542.6.2.269

Bird, B. M., Jofre, V. S. C., Geniole, S. N., Welker, K. M., Zilioli, S., Maestripieri, D., Arnocky, S., & Carré, J. M. (2016). Does the facial width-to-height ratio map onto variability in men's testosterone concentrations? *Evolution and Human Behavior*, *37*(5), 392-398. https://doi.org/10.1016/j.evolhumbehav.2016.03.004

Carré, J. M., & McCormick, C. M. (2008). In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proc Biol Sci*, *275*(1651), 2651-2656. https://doi.org/10.1098/rspb.2008.0873

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

Chatterjee, A., & Vartanian, O. (2016). Neuroscience of aesthetics. *Ann N Y Acad Sci*, *1369*(1), 172-194. https://doi.org/10.1111/nyas.13035

Costa, M., Lio, G., Gomez, A., & Sirigu, A. (2017). How components of facial width to height ratio differently contribute to the perception of social traits. *PLoS One*, *12*(2), e0172739. https://doi.org/10.1371/journal.pone.0172739

Delaney, H. D., & Maxwell, S. E. (1981). On using analysis of covariance in repeated measures designs. *Multivariate Behav Res*, *16*(1), 105-123. https://doi.org/10.1207/s15327906mbr1601_6

Deska, J. C., Lloyd, E. P., & Hugenberg, K. (2018a). 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

Deska, J. C., Lloyd, E. P., & Hugenberg, K. (2018b). Facing humanness: Facial width-to-height ratio predicts ascriptions of humanity. *Journal of Personality and Social Psychology*, *114*(1), 75-94. https://doi.org/10.1037/pspi0000110

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. J. (2019). Shared Neural Mechanisms of Visual Perception and Imagery. *Trends Cogn Sci*, *23*(5), 423-434. https://doi.org/10.1016/j.tics.2019.02.004

Eisenbruch, A. B., Lukaszewski, A. W., Simmons, Z. L., Arai, S., & Roney, J. R. (2018). Why the wide face? Androgen receptor gene polymorphism does not predict men's facial width-to-height ratio. *Adaptive Human Behavior and Physiology*, *4*(2), 138-151. https://doi.org/10.1007/s40750-017-0084-x

Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues.* Malor Books.

Elshout, M., Nelissen, R. M. A., & van Beest, I. (2017). Your act is worse than mine: Perception bias in revenge situations. *Aggress Behav*, *43*(6), 553-557. https://doi.org/10.1002/ab.21713

Fan, X., Wang, F., Shao, H., Zhang, P., & He, S. (2020). The bottom-up and top-down processing of faces in the human occipitotemporal cortex. *Elife*, *9*. https://doi.org/10.7554/eLife.48764

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*, *39*(2), 175-191. https://doi.org/10.3758/bf03193146

Finke, R. A. (1990). *Creative Imagery: Discoveries and Inventions in Visualization.* Lawrence Erlbaum.

Fish, J. B. (2019). Response Art in Art Therapy: Historical and Contemporary Overview. *Art Therapy*, *36*(3), 10. https://doi.org/https://doi.org/10.1080/07421656.2019.1648915

Fuster, J. M. (2006). The cognit: a network model of cortical representation. *Int J Psychophysiol*, *60*(2), 125-132. https://doi.org/10.1016/j.ijpsycho.2005.12.015

Geniole, S. N., Denson, T. F., Dixson, B. J., Carré, J. M., & McCormick, C. M. (2015). Evidence from meta-analyses of the facial width-to-height ratio as an evolved cue of threat. *PLoS One*, *10*(7), e0132726. https://doi.org/10.1371/journal.pone.0132726

Geniole, S. N., Keyes, A. E., Carré, J. M., & McCormick, C. M. (2014). Fearless dominance mediates the relationship between the facial width-to-height ratio and willingness to cheat. *Personality and Individual Differences*, *57*, 59-64. https://doi.org/10.1016/j.paid.2013.09.023

Goetz, S. M. M., Shattuck, K. S., Miller, R. M., Campbell, J. A., Lozoya, E., Weisfeld, G. E., & Carré, J. M. (2013). Social status moderates the relationship between facial structure and aggression. *Psychological Science*, *24*(11), 2329-2334. https://doi.org/10.1177/0956797613493294

Gu, S., Liu, Y., Liang, F., Feng, R., Li, Y., Liu, G., Gao, M., Liu, W., Wang, F., & Huang, J. H. (2020). Screening Depressive Disorders With Tree-Drawing Test. *Front Psychol*, *11*, 1446. https://doi.org/10.3389/fpsyg.2020.01446

Hahn, T., Winter, N. R., Anderl, C., Notebaert, K., Wuttke, A. M., Clement, C. C., & Windmann, S. (2017). Facial width-to-height ratio differs by social rank across organizations, countries, and value systems. *PLoS One*, *12*(11). https://doi.org/ARTN e0187957

Harms, C., & Lakens, D. (2018). Making 'null effects' informative: statistical techniques and inferential frameworks. *J Clin Transl Res*, *3*(Suppl 2), 382-393. https://www.ncbi.nlm.nih.gov/pubmed/30873486

Haselhuhn, M. P., Ormiston, M. E., & Wong, E. M. (2015). Men's facial width-to-height ratio predicts aggression: A meta-analysis. *PLoS One*, *10*(4). https://doi.org/10.1371/journal.pone.0122637

Haselhuhn, M. P., & Wong, E. M. (2012). Bad to the bone: facial structure predicts unethical behaviour. *Proc Biol Sci*, *279*(1728), 571-576. https://doi.org/10.1098/rspb.2011.1193

Haselhuhn, M. P., Wong, E. M., & Ormiston, M. E. (2013). Self-fulfilling prophecies as a link between men's facial width-to-height ratio and behavior. *PLoS One*, *8*(8). https://doi.org/ARTN e72259

He, X. J., Yin, H. F., Zeng, Y. C., Zhang, H., & Zhao, H. L. (2019). Facial structure and achievement drive: Evidence from financial analysts. *Journal of Accounting Research*, *57*(4), 1013-1057. https://doi.org/10.1111/1475-679x.12259

Hehman, E., Flake, J. K., & Freeman, J. B. (2015). Static and dynamic facial cues differentially affect the consistency of social evaluations. *Pers Soc Psychol Bull*, *41*(8), 1123-1134. https://doi.org/10.1177/0146167215591495

Hehman, E., Leitner, J. B., Deegan, M. P., & Gaertner, S. L. (2013). Facial structure is indicative of explicit support for prejudicial beliefs. *Psychol Sci*, *24*(3), 289-296. https://doi.org/10.1177/0956797612451467

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019). *Who judges a book by its cover? The prevalence, structure, and correlates of lay beliefs in physiognomy.* https://doi.org/10.31234/osf.io/8dq4x

Jeffreys, H. (1961). *Theory of Probability* (3 ed.). Oxford University Press.

Kamiya, S., Kim, Y. H., & Park, S. (2019). The face of risk: CEO facial masculinity and firm risk. *European Financial Management*, *25*(2), 239-270. https://doi.org/10.1111/eufm.12175

Kordsmeyer, T. L., Freund, D., Pita, S. R., Junger, J., & Penke, L. (2019). Further evidence that facial width-to-height ratio and global facial masculinity are not positively associated with testosterone levels. *Adaptive Human Behavior and Physiology*, *5*(2), 117-130. https://doi.org/10.1007/s40750-018-0105-4

Kramer, R. S. S. (2017). Sexual dimorphism of facial width-to-height ratio in human skulls and faces: A meta-analytical approach. *Evolution and Human Behavior*, *38*(3), 414-420. https://doi.org/10.1016/j.evolhumbehav.2016.12.002

Krenn, B., & Meier, J. (2018). Does facial width-to-height ratio predict aggressive behavior in association football? *Evolutionary Psychology*, *16*(4). https://doi.org/10.1177/1474704918818590

Krumhuber, E., Manstead, A. S., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, *7*(4), 730-735. https://doi.org/10.1037/1528-3542.7.4.730

LaBar, K. S., Crupain, M. J., Voyvodic, J. T., & McCarthy, G. (2003). Dynamic perception of facial affect and identity in the human brain. *Cereb Cortex*, *13*(10), 1023-1033. https://doi.org/10.1093/cercor/13.10.1023

LeDoux, J., & Daw, N. D. (2018). Surviving threats: neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nat Rev Neurosci*, *19*(5), 269-282. https://doi.org/10.1038/nrn.2018.22

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Lefevre, C. E., Etchells, P. J., Howell, E. C., Clark, A. P., & Penton-Voak, I. S. (2014). Facial width-to-height ratio predicts self-reported dominance and aggression in males and females, but a measure of masculinity does not. *Biology Letters*, *10*(10), 20140729. https://doi.org/10.1098/rsbl.2014.0729

Lefevre, C. E., & Lewis, G. J. (2014). Perceiving aggression from facial structure: Further evidence for a positive association with Facial width-to-height ratio and masculinity, but not for moderation by self-reported dominance. *European Journal of Personality*, *28*(6), 530-537. https://doi.org/10.1002/per.1942

Lefevre, C. E., Lewis, G. J., Perrett, D. I., & Penke, L. (2013). Telling facial metrics: facial width is associated with testosterone levels in men. *Evolution and Human Behavior*, *34*(4), 273-279. https://doi.org/10.1016/j.evolhumbehav.2013.03.005

Lewis, G. J., Lefevre, C. E., & Bates, T. C. (2012). Facial width-to-height ratio predicts achievement drive in US presidents. *Personality and Individual Differences*, *52*(7), 855-857. https://doi.org/10.1016/j.paid.2011.12.030

Lieberz, K. A., Muller-Engelmann, M., Bornefeld-Ettmann, P., Priebe, K., Weidmann, A., Fydrich, T., Geniole, S. N., McCormick, C. M., Rausch, S., Thome, J., & Steil, R. (2018). Detecting implicit cues of aggressiveness in male faces in revictimized female PTSD patients and healthy controls. *Psychiatry Research*, *267*, 429-437. https://doi.org/10.1016/j.psychres.2018.05.061

Lieberz, K. A., Windmann, S., Geniole, S. N., McCormick, C. M., Mueller-Engelmann, M., Gruener, F., Bornefeld-Ettmann, P., & Steil, R. (2017). The facial width-to-height ratio determines interpersonal distance preferences in the observer. *Aggressive Behavior*, *43*(5), 460-470. https://doi.org/10.1002/ab.21704

Lin, C., Adolphs, R., & Alvarez, R. M. (2018). Inferring whether officials are corruptible From looking at their faces. *Psychol Sci*, *29*(11), 1807-1823. https://doi.org/10.1177/0956797618788882

Loeb, E. L., Hessel, E. T., & Allen, J. P. (2016). The self-fulfilling prophecy of adolescent social expectations. *Int J Behav Dev*, *40*(6), 555-564. https://doi.org/10.1177/0165025415618274

MacDonell, E. T., Geniole, S. N., & McCormick, C. M. (2018). Force versus fury: Sex differences in the relationships among physical and psychological threat potential, the facial width-to-height ratio, and judgements of aggressiveness. *Aggress Behav*. https://doi.org/10.1002/ab.21771

Marsh, A. A., Cardinale, E. M., Chentsova-Dutton, Y. E., Grossman, M. R., & Krumpos, K. A. (2014). Power plays: Expressive mimicry of valid agonistic cues. *Social Psychological and Personality Science*, *5*(6), 7. https://doi.org/10.1177/1948550613519684

Martin, J. S., Staes, N., Weiss, A., Stevens, J. M. G., & Jaeggi, A. V. (2019). Facial width-to-height ratio is associated with agonistic and affiliative dominance in bonobos (Pan paniscus). *Biol Lett*, *15*(8), 20190232. https://doi.org/10.1098/rsbl.2019.0232

Matsumoto, D., & Hwang, H. C. (2021). Facial width-to-height ratios and deception skill. *Personality and Individual Differences*, *174*, 110683. https://doi.org/10.1016/j.paid.2021.110683.

Miall, R. C., Gowen, E., & Tchalenko, J. (2009). Drawing cartoon faces – a functional imaging study of the cognitive neuroscience of drawing. *Cortex*, *45*(3), 394-406. https://doi.org/10.1016/j.cortex.2007.10.013

Michelon, P., & Koenig, O. (2002). On the relationship between visual imagery and visual perception: evidence from priming studies. *European Journal of Cognitive Psychology*, *14*(2), 23. https://doi.org/https://doi.org/10.1080/09541440143000014

Mills, J., & Hogan, K. M. (2020). CEO facial masculinity and firm financial outcomes. *Corporate Board: Role, Duties and Composition, 16*(1), 39-46. https://doi.org/10.22495/cbv16i1art4

Mukunda, N., Moghbeli, N., Rizzo, A., Niepold, S., Bassett, B., & DeLisser, H. M. (2019). Visual art instruction in medical education: a narrative review. *Med Educ Online*, *24*(1), 1558657. https://doi.org/10.1080/10872981.2018.1558657

Noser, E., Schoch, J., & Ehlert, U. (2018). The influence of income and testosterone on the validity of facial width-to-height ratio as a biomarker for dominance. *PLoS One*, *13*(11). https://doi.org/10.1371/journal.pone.0207333

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

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

Sayõl, M. (2001). Children's drawings of emotional faces. *British Journal of Developmental Psychology*, *19*(4), 493-505. https://doi.org/10.1348/026151001166218

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspect Psychol Sci*, *12*(6), 1123-1128. https://doi.org/10.1177/1745691617708630

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *J Exp Psychol Gen*, *117*(1), 34-50. https://doi.org/10.1037//0096-3445.117.1.34

Stier, P. (2019). *"Es stand ihm ins Gesicht geschrieben" - Der Einfluss stereotyper Attribution auf den Zusammenhang zwischen Gesichtsmorphologie und Persönlichkeit* [Master Thesis, Johann-Wolfgang-Goethe University Frankfurt]. https://doi.org/10.17605/osf.io/mtzax

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

Terada, K., & Takeuchi, C. (2017). Emotional Expression in Simple Line Drawings of a Robot's Face Leads to Higher Offers in the Ultimatum Game. *Front Psychol*, *8*, 724. https://doi.org/10.3389/fpsyg.2017.00724

Tipples, J., Atkinson, A. P., & Young, A. W. (2002). The eyebrow frown: a salient social signal. *Emotion*, *2*(3), 288-296. https://doi.org/10.1037/1528-3542.2.3.288

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annu Rev Psychol*, *66*, 519-545. https://doi.org/10.1146/annurev-psych-113011-143831

Trautmann, S. A., Fehr, T., & Herrmann, M. (2009). Emotions in motion: dynamic compared to static facial expressions of disgust and happiness reveal more widespread emotion-specific activations. *Brain Res*, *1284*, 100-115. https://doi.org/10.1016/j.brainres.2009.05.075

Trebicky, V., Fialova, J., Kleisner, K., & Havlicek, J. (2016). Focal length affects depicted shape and perception of facial images. *PLoS One*, *11*(2). https://doi.org/10.1371/journal.pone.0149313

Trebicky, V., Fialova, J., Kleisner, K., Roberts, S. C., Little, A. C., & Havlicek, J. (2015). Further evidence for links between facial width-to-height ratio and fighting success: Commentary on Zilioli et al. (2014). *Aggress Behav*, *41*(4), 331-334. https://doi.org/10.1002/ab.21559

Tsujimura, H., & Banissy, M. J. (2013). Human face structure correlates with professional baseball performance: insights from professional Japanese baseball players. *Biol Lett*, *9*(3), 20130140. https://doi.org/10.1098/rsbl.2013.0140

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. *Psychol Sci*, *25*(3), 806-811. https://doi.org/10.1177/0956797613511823

Verstijnen, I. M., van Leeuwen, C., Goldschmidt, G., Hamel, R., & Hennessey, J. M. (1998). Creative discovery in imagery and perception: combining is relatively easy, restructuring takes a sketch. *Acta Psychol (Amst)*, *99*(2), 177-200. https://doi.org/10.1016/s0001-6918(98)00010-9

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E. J., van Doorn, J., Smira, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychon Bull Rev*, *25*(1), 58-76. https://doi.org/10.3758/s13423-017-1323-7

Wang, D. W., Nair, K., Kouchaki, M., Zajac, E. J., & Zhao, X. X. (2019). A case of evolutionary mismatch? Why facial width-to-height ratio may not predict behavioral tendencies. *Psychological Science*, *30*(7), 1074-1081. https://doi.org/10.1177/0956797619849928
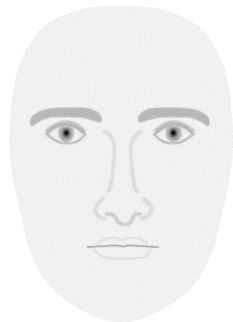
Welker, K. M., Goetz, S. M. M., Galicia, S., Liphardt, J., & Carré, J. M. (2015). An examination of the associations between facial structure, aggressive behavior, and performance in the 2010 World Cup Association football players. *Adaptive Human Behavior and Physiology*, *1*(1), 17-29. https://doi.org/10.1007/s40750-014-0003-3

Weston, E. M., Friday, A. E., & Lio, P. (2007). Biometric evidence that sexual selection has shaped the hominin face. *PLoS One*, *2*(8), e710. https://doi.org/10.1371/journal.pone.0000710

Weyers, P., Muhlberger, A., Hefele, C., & Pauli, P. (2006). Electromyographic responses to static and dynamic avatar emotional facial expressions. *Psychophysiology*, *43*(5), 450-453. https://doi.org/10.1111/j.1469-8986.2006.00451.x

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

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychol Sci*, *26*(8), 1325-1331. https://doi.org/10.1177/0956797615590992

Windmann, S., & Krüger, T. (1998). Subconscious detection of threat as reflected by an enhanced response bias. *Consciousness and Cognition, 7*(4), 603-633. https://doi.org/10.1006/ccog.1998.0337

Windmann, S., Steinbrück, L., & Stier, P. (2021). *Generating the face of a character* [Data, scripts, and materials]. Center for Open Science. https://doi.org/10.17605/osf.io/mtzax

Witkower, Z., & Tracy, J. L. (2020). How and why head position changes the perception of facial expressions of emotion. *Emotion*. https://doi.org/10.1037/emo0000846

Xie, S. Y., Flake, J. K., & Hehman, E. (2019). Perceiver and target characteristics contribute to impression formation differently across race and gender. *J Pers Soc Psychol*, *117*(2), 364-385. https://doi.org/10.1037/pspi0000160

Yalcinkaya, E., Cingi, C., Soken, H., Ulusoy, S., & Muluk, N. B. (2016). Aesthetic analysis of the ideal eyebrow shape and position. *Eur Arch Otorhinolaryngol*, *273*(2), 305-310. https://doi.org/10.1007/s00405-014-3356-0

Yang, Y., Tang, C., Qu, X., Wang, C., & Denson, T. F. (2018). Group facial width-to-height ratio predicts intergroup negotiation outcomes. *Front Psychol*, *9*, 214. https://doi.org/10.3389/fpsyg.2018.00214

Yildirim, B. O., & Derksen, J. J. (2012). A review on the relationship between testosterone and life-course persistent antisocial behavior. *Psychiatry Res*, *200*(2-3), 984-1010. https://doi.org/10.1016/j.psychres.2012.07.044

Zayas, V., Shoda, Y., & Ayduk, O. N. (2002). Personality in context: an interpersonal systems perspective. *J Pers*, *70*(6), 851-900. https://doi.org/10.1111/1467-6494.05026

Zebrowitz, L. A. (2017). First impressions from faces. *Curr Dir Psychol Sci*, *26*(3), 237-242. https://doi.org/10.1177/0963721416683996

Zebrowitz, L. A., Kikuchi, M., & Fellous, J. M. (2010). Facial resemblance to emotions: group differences, impression effects, and race stereotypes. *J Pers Soc Psychol*, *98*(2), 175-189. https://doi.org/10.1037/a0017990

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Soc Personal Psychol Compass*, *2*(3), 1497. https://doi.org/10.1111/j.1751-9004.2008.00109.x

Zhang, D., Lin, H., & Perrett, D. I. (2020). Apparent Emotional Expression Explains the Effects of Head Posture on Perceived Trustworthiness and Dominance, but a Measure of Facial Width Does Not. *Perception*, *49*(4), 422-438. https://doi.org/10.1177/0301006620909286
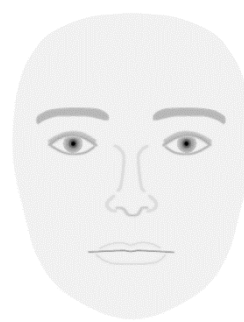
Appendix

Sample images from Study 1a

Aggressive-dominant	Peaceloving-submissive

Assembly

fWHR = 1.95   fWHR = 1.96

Photo-edited

fWHR = 1.94   fWHR = 1.52
state-emotionality = 4   state-emotionality = 7

Drawings

fWHR = 2.05   fWHR = 1.82
state-emotionality = 19   state-emotionality = 0

Note. Photograph-like images in the mid row have been generated using FaceGen© Artist software by Singular Inversions Incorporated, Toronto, Canada; reprinted with permission.