ARTICLE IN PRESS

Evolution and Human Behavior xxx (xxxx) xxx

Contents lists available at ScienceDirect

Evolution and Human Behavior

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

# Is facial width-to-height ratio reliably associated with social inferences?

Patrick K. Durkee [a,*], Jessica D. Ayers [b]

[a] Department of Psychology, The University of Texas at Austin, USA
[b] Department of Psychology, Arizona State University, USA

ARTICLE INFO

*Keywords:*
Person perception
Facial width-to-height ratio
Evolutionary mismatch
Byproduct

ABSTRACT

Theoretical considerations and early empirical findings suggested facial width-to-height ratio (fWHR) may be relevant to person perception because it is associated with behavioral dispositions. More recent evidence failing to find fWHR-behavior links suggests that mismatch or byproduct hypotheses may be necessary to explain fWHR-based trait inferences; however, these explanations may not be needed because it is not clear that fWHR is reliably associated with trait inferences. To investigate the robustness of fWHR-inference links, we conducted secondary analyses of a cross-national dataset consisting of ratings by 11,481 participants across 11 world regions who judged 60 male and 60 female faces on one of 13 social traits (*ns* per trait range from 760 to 975). In preregistered analyses—and exploratory analyses of a subset of traits in the larger sample of 597 faces from which the 120 faces were drawn—we found mixed evidence for fWHR-based social judgments. In multilevel models, fWHR was not reliably linked to raters' judgments of male faces for any of the 13 trait-inferences but was negatively associated with ratings of female faces' dominance, trustworthiness, sociability, emotional stability, responsibility, confidence, attractiveness, and intelligence. In exploratory analyses of a subset of traits using the larger sample of faces, fWHR was associated positively with perceptions of meanness and aggressiveness in male but not female faces, negatively with attractiveness and dominance in female but not male faces, and negatively with trustworthiness in male but not female faces. We interpret these mixed findings to suggest that (1) fWHR-inference links are likely to be smaller and less reliable than expected from prior research; (2) fWHR may play a larger role in perceptions of female faces than would be predicted from the theory underpinning fWHR hypotheses; and (3) future research should more closely examine the extent to which robust fWHR-inferences reflect mismatch in the reliability of fWHR-behavior links between ancestral and modern environments versus byproducts of other person perception mechanisms.

## 1. Introduction

Humans make consequential social judgments and inferences from faces (Todorov, Mende-Siedlecki, & Dotsch, 2013). Identifying the facial features on which these judgments are made is a major objective of the person-perception branch of psychological science (Lick & Johnson, 2018; Mason, Cloutier, & Macrae, 2006; Oosterhof & Todorov, 2008; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). Careful pruning of candidate features can help us better understand proximate and ultimate factors involved in human perception and social behavior.

Early research suggested that fWHR would be a promising candidate facial feature for deconstructing human perceptions and social behavior (e.g., Carré & McCormick, 2008; Carré, McCormick, & Mondloch, 2009; Geniole et al., 2012; Haselhuhn, Ormiston, & Wong, 2015). After more than a decade of research, however, the theoretical importance of fWHR in person perception is unclear. In this paper, we present findings from a cross-national dataset pertinent to determining the role of fWHR in person perception.

### 1.1. Brief overview of fWHR research

The fWHR metric was introduced by Weston, Friday, and Liò (2007). Their measurements of 121 dry human skulls (53 female) suggested that the morphology of faces diverge at puberty such that, in comparison to adult female faces, the bizygomatic width of adult male faces tend to be wider relative to facial height. Weston and colleagues interpreted this sexual dimorphism as evidence of sexual selection independent of selection for overall body size.

Based on this initial evidence, Carré and McCormick (2008) proposed that the development of fWHR may be related to other sexually-

* Corresponding author at: Department of Psychology, University of Texas at Austin, 108 E Dean Keeton St, Austin, TX 78712, USA.
*E-mail address:* pdurkee@utexas.edu (P.K. Durkee).

https://doi.org/10.1016/j.evolhumbehav.2021.06.003
Received 26 June 2020; Received in revised form 5 April 2021; Accepted 14 June 2021

dimorphic pubertal changes in hormones and behavior. They hypothesized that the organizational effects of testosterone, which drive sexual differentiation in adolescence, may be a common cause of within-sex variation in fWHR and aggressive behavior in men. In support of this hypothesis, they found that fWHR was positively associated with lab-based reactive aggression in undergraduate men ($n$ = 38) but not women ($n$ = 51), as well as in-game penalty minutes accrued by male hockey players at the college level ($n$ = 21) and professional level ($n$ = 112). Concurrent investigations found that undergraduate students' ($n$ = 47) face-based judgments of men's aggressiveness were reliably positively associated with men's fWHR and manifest aggressive behavior ($n$ = 37; Carré et al., 2009). These studies provided initial evidence that fWHR is linked to theoretically-relevant behavioral tendencies and judgments in men; further, they suggested that fWHR could be useful in examining social inferences and behavior more broadly.

Hundreds of studies have examined links between individual differences in fWHR, behavioral tendencies, and social inferences. Because the association between fWHR and behavioral tendencies was hypothesized to reflect common associations with developmental testosterone levels (Carré & McCormick, 2008), most of the research on the relationships between fWHR and behavior examines domains linked to testosterone, such as aggression. For instance, larger fWHR was reported to be associated with more fouls, goals, and assists in soccer players from 32 countries in the 2010 Men's World Cup (Welker et al., 2015); higher achievement drive in former US presidents (Lewis, Lefevre, & Bates, 2012); greater success in negotiation (Haselhuhn, Wong, Ormiston, Inesi, & Galinsky, 2014); as well as higher scores on Machiavellian traits (Noser et al., 2018). In line with these apparent behavioral associations, many studies suggested that fWHR may be a cue people use to make judgments about general behavioral dispositions. For example, people purportedly view men with larger fWHR as less trustworthy (e.g., Ormiston, Wong, & Haselhuhn, 2017; Stirrat & Perrett, 2010), less attractive (e.g., Boshyan, Zebrowitz, Franklin Jr, McCormick, & Carré, 2014; Stirrat & Perrett, 2010; Valentine, Li, Penke, & Perrett, 2014), more deceitful (e.g., Geniole, Keyes, Carré, & McCormick, 2014; Haselhuhn & Wong, 2011), more dominant (e.g., Mileva, Cowan, Cobey, Knowles, & Little, 2014; Valentine et al., 2014), and more aggressive (e.g., Geniole & McCormick, 2015; Lefevre & Lewis, 2014). Meta-analyses of the early fWHR literature suggested that the associations between fWHR, behavioral tendencies, and social perceptions across studies were small (behavioral tendencies) to moderate (social perceptions) in size (Geniole, Denson, Dixson, Carré, & McCormick, 2015; Haselhuhn et al., 2015).

However, there are several reasons to question the theoretical core of fWHR research and the ostensible associations between fWHR, behavioral tendencies, and social perceptions. First, while Carré and McCormick (2008) found evidence of sexual dimorphism in fWHR from facial photographs mirroring Weston et al. (2007), subsequent studies suggest that fWHR is not reliably sexually dimorphic in humans (Kramer, 2017; Kramer et al., 2012; Lefevre et al., 2012; Özener, 2012; Robertson & Kingsley, 2018; Stirrat, Stulp, & Pollet, 2012; but see Köllner, Janson, & Schultheiss, 2018). Second, although early research found some support for hypothesized links between testosterone and fWHR (Lefevre, Lewis, Perrett, & Penke, 2013), large-scale studies and meta-analyses using a broad range of methodologies fail to support associations between testosterone and fWHR (Bird et al., 2016; Eisenbruch, Lukaszewski, Simmons, Arai, & Roney, 2018; Hodges-Simeon, Sobraske, Samore, Gurven, & Gaulin, 2016; Kordsmeyer, Freund, Pita, Jünger, & Penke, 2019). Third, previous studies examining behavioral links tended to examine lab-based behaviors with relatively small sample sizes, but recent large-scale replications of links between fWHR and behavioral tendencies found no evidence for theoretically-predicted behavioral links (Kosinski, 2017; Kramer, 2015; Wang, Nair, Kouchaki, Zajac, & Zhao, 2019). Although several studies and a meta-analysis based largely on Western samples found some evidence that fWHR may be linked to social inferences (Geniole et al., 2015), no large-scale, cross-cultural studies have examined these effects. Because meta-analyses are highly susceptible to publication bias and can grossly overestimate effect sizes (Kvarven, Strømland, & Johannesson, 2019), it is crucial to reexamine the reliability of associations between fWHR and trait inferences.

Importantly, even if fWHR is not sexually dimorphic, not influenced by testosterone, nor associated with behavioral tendencies, trait judgments could still be influenced by fWHR. For instance, associations between fWHR and social inferences could reflect evolutionary mismatch: the phenomenon whereby some trait or mechanism that was adaptive in ancestral environments is no longer adaptive—or even maladaptive—in current environments (c.f., Frankenhuis & Del Giudice, 2012; Li, van Vugt, & Colarelli, 2018). Alternatively, links between social inferences and variation in fWHR could be a byproduct of other adaptative mechanisms. We discuss each of these possibilities in more detail below.

Inferences based on fWHR would reflect an evolutionary mismatch if the mind possesses mechanisms designed to use fWHR to make social inferences because fWHR was probabilistically associated with behavioral dispositions throughout human evolution but is no longer associated with behavior in contemporary human environments. For example, within-sex variation in testosterone could have been a common cause of variation in fWHR and aggression in the ancestral past, making it potentially adaptive at that time for humans to partially base inferences of others on fWHR in order to avoid aggressive individuals. If this were the case, selective pressures could have slowly tuned social-perception mechanisms within the mind to base inferences of likely behaviors—such as the propensity of an individual with a high fWHR to be aggressive—off of this static facial cue. In more recent history, however, social pressures and sanctions that are associated with reductions in manifest aggressive behavior (c.f., Pinker, 2012) may truncate any ancestrally-reliable links between fWHR and actual aggressive behaviors in modern environments. But if the mind were already designed by evolutionary processes to interpret fWHR as a probabilistic indicator of others' aggressiveness, these inferential biases would remain even in the absence of links between fWHR and behavior in the modern world.

It is also possible that the link between fWHR and trait-inferences is a byproduct of other functional person-perception mechanisms. An alternative, but non-mutually exclusive, explanation is that associations between fWHR and trait-inferences could be a byproduct if they result from mechanisms that are not designed specifically to infer behavioral tendencies from fWHR. For example, perceptions of fWHR could be a byproduct of anger recognition: faces with larger fWHRs are more likely to be rated as angry when neutral compared to faces with smaller fWHRs (Deska, Almaraz, & Hugenberg, 2017; Neth & Martinez, 2010), and anger detection mechanisms may infer a host of likely behavioral outputs of that perceived anger (e.g., aggression, defection) that then biases ratings of the targets dispositions. Additionally, since fWHR decreases with age, trait-inferences from fWHR may be a byproduct of age estimation mechanisms (Hehman, Leitner, & Freeman, 2014; Robertson, Kingsley, & Ford, 2017). For example, higher fWHRs may be interpreted as a cue to youth, and therefore, higher perceived likelihood of aggression in males (Archer, 2019); or, in female faces, age inferences could cue potential reproductive value driving links between attractiveness ratings (Lassek & Gaulin, 2019).

### 1.2. The current study

Of course, mismatch or byproduct explanations for manifest associations between social inferences and fWHR would only be necessary if the human mind reliably makes fWHR-based social inferences. Although fWHR may not be sexually dimorphic, influenced by testosterone, or reliably linked with behavioral tendencies, an open question in the fWHR literature remains: Is facial width-to-height ratio reliably associated with social inferences? The answer to this question would help to determine whether further research investigating evolutionary mismatch and byproduct explanations is needed. To inform this outstanding question, we leveraged a large cross-national dataset assembled by Jones et al. (2021). The dataset contains ratings of 120 faces by raters across 41 countries and 11 world regions on 13 fundamental social traits: aggressive, attractive, caring, confident, dominant, emotionally stable, intelligent, mean, responsible, sociable, trustworthy, unhappy, and weird (Oosterhof & Todorov, 2008).

Several predictions follow logically from the theoretical considerations and seminal research underpinning the fWHR research program. First, in male faces, fWHR should be positively associated with perceptions of aggressiveness and dominance (Geniole et al., 2015). Second, in male faces, fWHR should be negatively associated with perceptions of trustworthiness and attractiveness (Geniole et al., 2015; Stirrat & Perrett, 2010). These associations are predicted to be attenuated or absent in women (Geniole et al., 2015). No firm predictions can be derived from the literature about the associations between fWHR and judgments of how caring, confident, emotionally stable, intelligent, mean, responsible, sociable, unhappy, or weird men and women look. Still, we examined associations between fWHR and these trait judgments because they may be informative in evaluating the discriminant validity and general parameters of fWHR-based social judgments. We did not preregister specific predictions for any of the 13 traits, but we preregistered our analyses using an exploratory subset of the Jones et al. (2021) data prior to the release of the full dataset (https://osf.io/ykh4c/), which include a description of the research question and implies the hypotheses that will be tested.

## 2. Methods

### 2.1. Participants

The full dataset, provided by Jones et al. (2021) as part of the Psychological Science Accelerator Secondary Analysis Challenge,[1] consists of ratings from 11,481 raters (7967 women) across 44 countries, representing 11 world regions and 25 languages. Country-level sample sizes range from 27 to 2273 ($M = 260.93$, $Mdn = 166.50$, $SD = 353.85$). Figure 1 highlights the relative sample sizes from each nation. The mean age of participants is 22.56 ($Mdn = 20$, $SD = 6.97$).

### 2.2. Study materials and procedure

#### 2.2.1. Face stimuli and fWHR measurement

The face stimuli rated in the study are taken from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015). All subjects displayed a neutral expression, a straight-ahead gaze, and wore a gray t-shirt; the full-color photos were taken under standardized lighting with a uniform background. The stimuli set selected for rating in Jones et al. (2021) consists of a subset of 60 male and 60 female faces, with equal numbers of Asian, Black, Latino, and White faces—30 faces each, with 15 male and 15 female faces per group. The ages of subjects range from 18 to 35 ($M = 26.38$, $SD = 3.57$).

In our preregistered analyses we used fWHR measurements provided with the Chicago Face Database norming data. However, upon further inspection it became clear that these fWHR scores were calculated using the distance between cheekbones to measure face width rather than the more standard bizygomatic width.[2] This differently-calculated fWHR may not map onto the common theoretical operationalization of fWHR. We therefore calculated new fWHR measurements for the 120 faces that were rated as part of the Jones, Schild, and Jones (2020) study based on more common operationalization of the bizygomatic width of the face divided by the distance between the upper lip and the ridge of the eyebrow (Lefevre et al., 2013). Three research assistants manually measured the bizygomatic width and height using the software ImageJ and procedures outlined in Lefevre et al. (2013). The three sets of manually measured fWHRs were all highly intercorrelated with one another (*r*s from 0.92–0.96), so we took an average across the three manual measurements for each face.

Additionally, we used a recently-published automated method to calculate fWHR (Jones et al., 2020). Because the automated procedure has only been validated on two sets of faces, we compared the fWHR scores obtained from the manual measurements to the scores provided by the automated procedure, so that we could further validate the automated method and examine any potential ethnicity bias in the automated procedure.

The fWHR score from the manual measurements was highly correlated with the fWHR score from the automated procedure ($r = 0.93$). Because Jones et al. (2020) found some evidence that the automated measure was more accurate for East Asian faces than White faces, we examined whether there was bias in the automated measurements such that the correlation between manual measures and the automated measures differed as a function of the face sex or ethnicity. The face ethnicity and sex specific correlations are depicted in Fig. 2. We found no statistically significant ethnicity, sex, or ethnicity by sex interactions; however, these tests may be somewhat underpowered given the small sample sizes when grouped by ethnicity and sex. For the sake of comparison, we report our confirmatory results based on both the manual and the automated fWHR measurements.

#### 2.2.2. Trait ratings

Rating data were collected by 126 labs involved with the Psychological Science Accelerator. All data collection sites obtained approval from their local IRB, unless the task was covered by a pre-existing IRB approval or this type of rating task was classified as exempt by the local IRB.

After providing informed consent and completing a demographic questionnaire, participants were randomly assigned to rate each of the 120 faces on one of 13 trait adjectives. The 13 trait adjectives—aggressive, attractive, caring, confident, dominant, emotionally stable, intelligent, mean, responsible, sociable, trustworthy, unhappy, and weird—were taken from Oosterhof and Todorov (2008). Participants rated the degree to which the adjective applied to each face using a 9-point scale (1 = *not at all*; 9 = *very*). Instructions and rating prompts were presented to each participant in the appropriate language for their site. The same face stimuli were used in each testing site. Faces were presented in a randomized order for each participant. Participants completed the ratings twice and these ratings were averaged for each person prior to the dataset release. Table 1 shows the number of raters for each trait, along with the mean and standard deviation of each trait rating across all 120 faces.

## 3. Results

We conducted all data tidying and analyses in *R*. All code used to conduct the analyses and create figures presented is publicly available on our associated Open Science Framework project page (https://osf.io/ykh4c/). The full Jones et al. (2021) dataset used in our confirmatory analyses is also on the Open Science Framework (https://osf.io/jfwtr/).

### 3.1. Confirmatory analyses

We preregistered our analyses based on an exploratory data set consisting of a random subset of one-third of raters from each lab ($n = 3851$), which was released by the Psychological Science Accelerator as part of the Secondary Analysis Challenge. We ran separate multilevel models for ratings of male and female faces (henceforth *targets*) because of theory-driven expectations for sex-differentiated associations (e.g., Carré & McCormick, 2008). Each model examined one of the 13 trait-rating associations with fWHR, regressing participants' scaled ratings

[1] Information about this challenge can be found at https://psysciacc.org/2019/09/01/introducing-the-psa001-secondary-analysis-challenge/

[2] We thank an anonymous reviewer for pointing this out.

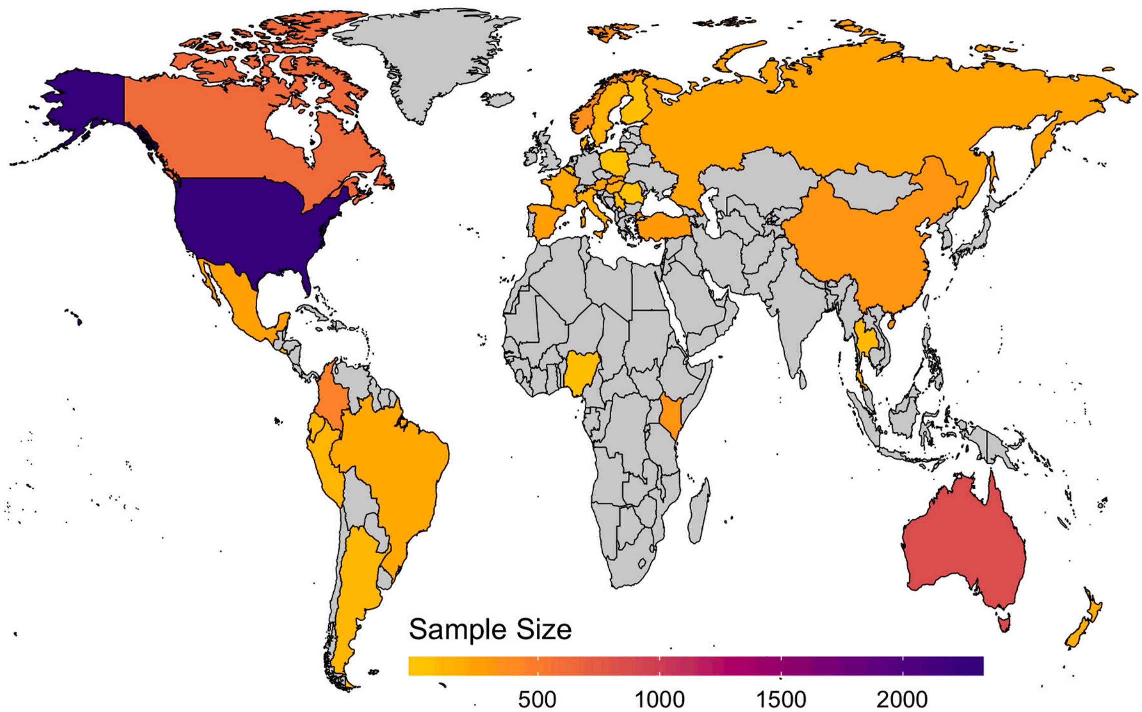


**Fig. 1.** World map highlighting the sample sizes from different world regions in the full Psychological Science Accelerator dataset.

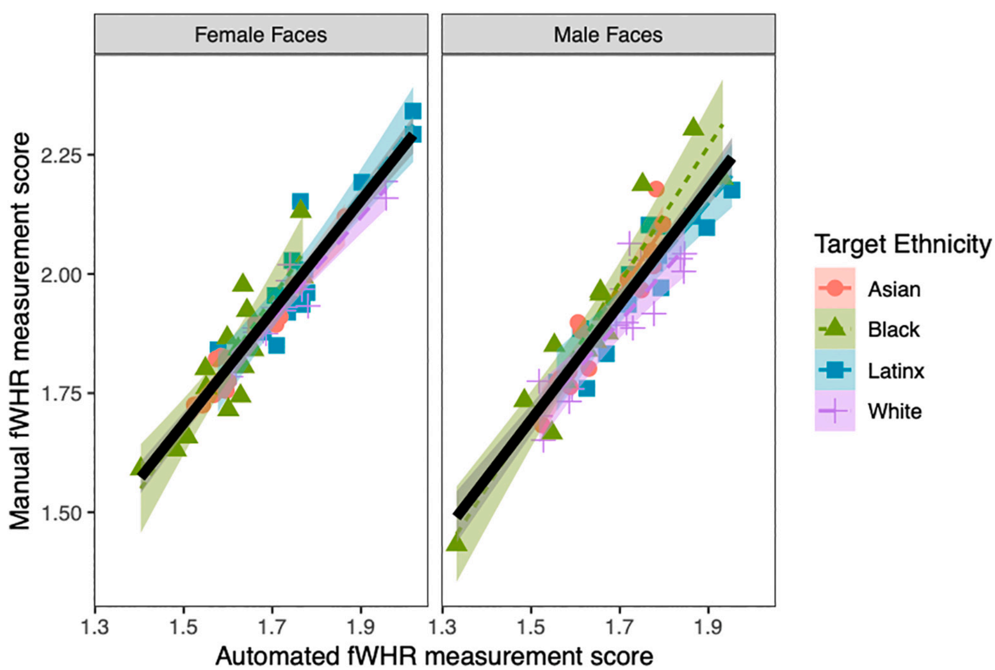


**Fig. 2.** Scatterplots depicting the target sex and ethnicity specific correlations between fWHR scores calculated using manual measurement methods and automated measurement methods.

for a given trait on targets' fWHR, resulting in 26 total models (13 trait models for male targets and 13 trait models for female targets).

In the initial preregistered analyses, we specified random intercepts for targets with random slopes and random intercepts for raters, labs, countries, and world regions to account for all levels of nesting within the data. After reproducibility checks conducted by the Psychological Science Accelerator in January 2020, we decided to remove the random intercept and slope specifications for labs, countries, and world regions because the effectively zero variability at these higher-order levels caused convergence issues and often triggered singular fit warnings.[3] We preregistered these simplified models to conduct the confirmatory analyses prior to the release of the confirmatory dataset.

The random-effects structures in the final preregistered analyses

[3] We thank Patrick Forscher for conducting reproducibility checks on our exploratory analyses and providing suggestions for resolving convergence issues prior to confirmatory analyses.

**Table 1**
Number of raters rating each trait and descriptive statistics for each trait rating.

| Trait rating | *n* | *M* | *SD* |
|---|---|---|---|
| Aggressive | 854 | 3.96 | 1.83 |
| Attractive | 832 | 3.97 | 1.81 |
| Caring | 906 | 4.71 | 1.70 |
| Confident | 975 | 5.07 | 1.63 |
| Dominant | 896 | 4.66 | 1.72 |
| Emotional Stability | 896 | 5.43 | 1.60 |
| Intelligent | 895 | 5.41 | 1.56 |
| Mean | 839 | 3.96 | 1.81 |
| Responsible | 880 | 5.57 | 1.64 |
| Sociable | 958 | 4.88 | 1.60 |
| Trustworthy | 896 | 5.16 | 1.64 |
| Unhappy | 897 | 4.40 | 1.75 |
| Weird | 760 | 3.87 | 2.03 |

Note: Faces were rated on a 9-point scale (1 = *not at all*; 9 = *very*). *SD* = Standard Deviation; *M* = Mean; *n* = number of raters for trait.

were specified as follows. For all models, we specified random intercepts for targets which allows us to generalize to the wider population of human faces represented by the faces in the face set (Judd, Westfall, & Kenny, 2012). For most models, we specified random intercepts and slopes for raters, allowing for generalization to the greater population of raters, represented by the samples of raters who participated in the study. In the preregistration stage, we removed the random slopes for raters in four models—male attractiveness, male unhappiness, male emotional stability, and male sociability—because of convergence issues with these slope specifications (c.f., Barr, Levy, Scheepers, & Tily, 2013). In the final confirmatory analyses, two more models—male caring and male confidence—failed to converge, so we removed the random slopes specification for raters in these models. We adopted the conventional alpha level of 0.05 to assess statistical significance of associations.

Figure 3 shows the point estimate, standard error, and statistical significance level for each sex-specific relationship between fWHR and the 13 social adjectives using fWHR scores based on the manual measurement (Geniole et al., 2013) and automated method (Jones et al., 2020). The results were not qualitatively different between the different fWHR measurement methods. For male faces, only the relationship between fWHR and intelligence ratings was statistically significant in under both measurement methods (*ps* < 0.047), while no other relationships between fWHR and the 12 trait ratings were statistically significant using fWHR scores from either method (0.123 < *ps* < 0.982; $M_p$ = 0.374). For female faces, we found small but statistically significant negative associations between female fWHR and rated dominance, trustworthiness, sociability, emotional stability, responsibility, confidence, attractiveness, and intelligence, (0.0001 < *ps* < 0.041; $M_p$ = 0.017); associations between fWHR and ratings of how unhappy, aggressive, mean, weird, and caring the female faces appeared were not statistically significant in either the exploratory or confirmatory sample (0.051 < *ps* < 0.424; $M_p$ = 0.184). Regression tables for each trait- and sex-specific analysis, as well as plots depicting the country-level and participant-level slopes for each of the 13 sex-specific trait associations for both fWHR measurement methods are provided on the OSF (https://osf.io/ykh4c/).

### 3.2. Exploratory analyses

The 120 faces used in the Jones et al. (2021) study that we have used in the current study are only a subset of the larger set of 597 faces available in the CFD photoset, which contains photos of 57 Asian females, 53 Asian males, 95 Black females, 108 Black males, 56 Latinas, 52 Latinos, 90 White females, and 94 White males (Ma et al., 2015). We sought to address concerns about statistical power and representativeness afforded by the subset of 120 CFD faces rated in the PSA study by comparing relevant associations to the full sample. Second, we explored differences in fWHR-perception links as a function of target ethnicity.

Because we did not design the study from which we drew our data (i.e., Jones et al., 2021), we can only provide post-hoc power analyses. The design of the PSA study from which we obtained our data closely matches a stimuli-within-condition design (c.f., Westfall et al., 2014). Based on the power curves for such a design published in Westfall et al. (2014), our multilevel models with 60 faces per sex and at least 700 raters per trait may only provide around 78% power to detect effects as small as *r* = *0*.20 (*d* = *0*.4). It is important to note, however, that power in multilevel models depends on many factors, such as the relative degree of the stimulus and rater intercept variance, rater slope variance, and residual variance. In a post-hoc power analysis using the online app[4] associated with Westfall et al. (2014) that specified variance component patterns that resemble what we observed in our multilevel models, such that the stimulus intercept and the rater slope variance components are small relative to the rater intercept and residual variance components, it appears we have around 80% power to detect effects as small as *d* = *0*.34 (*r* = 0.17) and around 92% power to detect the lower-end meta-analysis estimates of *r* = 0.2 (*d* = 0.4). Thus, our confirmatory analyses appear to have been adequately powered to detect associations between fWHR and trait ratings that reflect the low-end estimates of *r* = 0.20 found in meta-analyses of fWHR threat-perception links (Geniole et al., 2015), and certainly the mean effect size estimate of *r* = 0.46.

Still, the 120 faces used in the PSA rating study are only a subset of the 597 faces (307 female) that are available in the CFD photoset (Ma et al., 2015). It is possible that they are not representative of the larger CFD sample. The CFD norming data contains averaged trait ratings on several traits—angry, attractive, dominant, trustworthy, threatening—that conceptually overlap with the traits rated by the raters in the PSA study—mean, aggressive, dominant, attractive, trustworthy. We therefore sought to examine the extent to which the associations based on the subset of faces in the PSA are different from the associations in the full CFD target sample. However, the CFD ratings for each target are based on smaller samples of raters that are not as diverse as the participants who provided trait-ratings for the PSA study (Ma et al., 2015); they cannot be generalized to new samples of raters from around the world as ratings from the PSA study can. Because the CFD norming data contains only averaged ratings for each target, rather than the individual ratings, we cannot use multilevel models to directly compare the associations in both the subsample and the full target sample. For comparison between the PSA ratings and the CFD ratings, we therefore computed each face's average PSA rating across world regions for each trait to match the structure of the CFD norming data.

Importantly, the subset of CFD faces chosen to be rated in the PSA were chosen at random (Jones et al., 2021), so whether a given face was included in the PSA ratings has no association with their fWHR. The missing rating data for the remainder of the faces can therefore be considered missing completely at random (MCAR). We can therefore use Full Information Maximum Likelihood Estimation (FIML) to account for the missing PSA ratings and obtain unbiased coefficient estimates of the fWHR-perception associations across all targets (Enders, 2001; Enders & Bandalos, 2001). In this case, FIML uses the covariances between the conceptually-similar CFD trait-ratings, averaged PSA trait-ratings, and fWHR scores to inform the estimates of the associations between averaged PSA trait-ratings and fWHR across all 597 faces. Additionally, in order to factor in variability of associations that would result from differences in average trait-ratings derived from different samples of raters, different samples of faces, and the resulting trait correlations that may influence the FIML estimates, we utilized a bootstrapping procedure where the following steps were repeated 5000 times:

1) a new sample of PSA ratings were selected with replacement from the original PSA dataset, such that the number of trait-ratings selected

[4] The power analysis with crossed designs app is available at http://jakewestfall.org/power/.

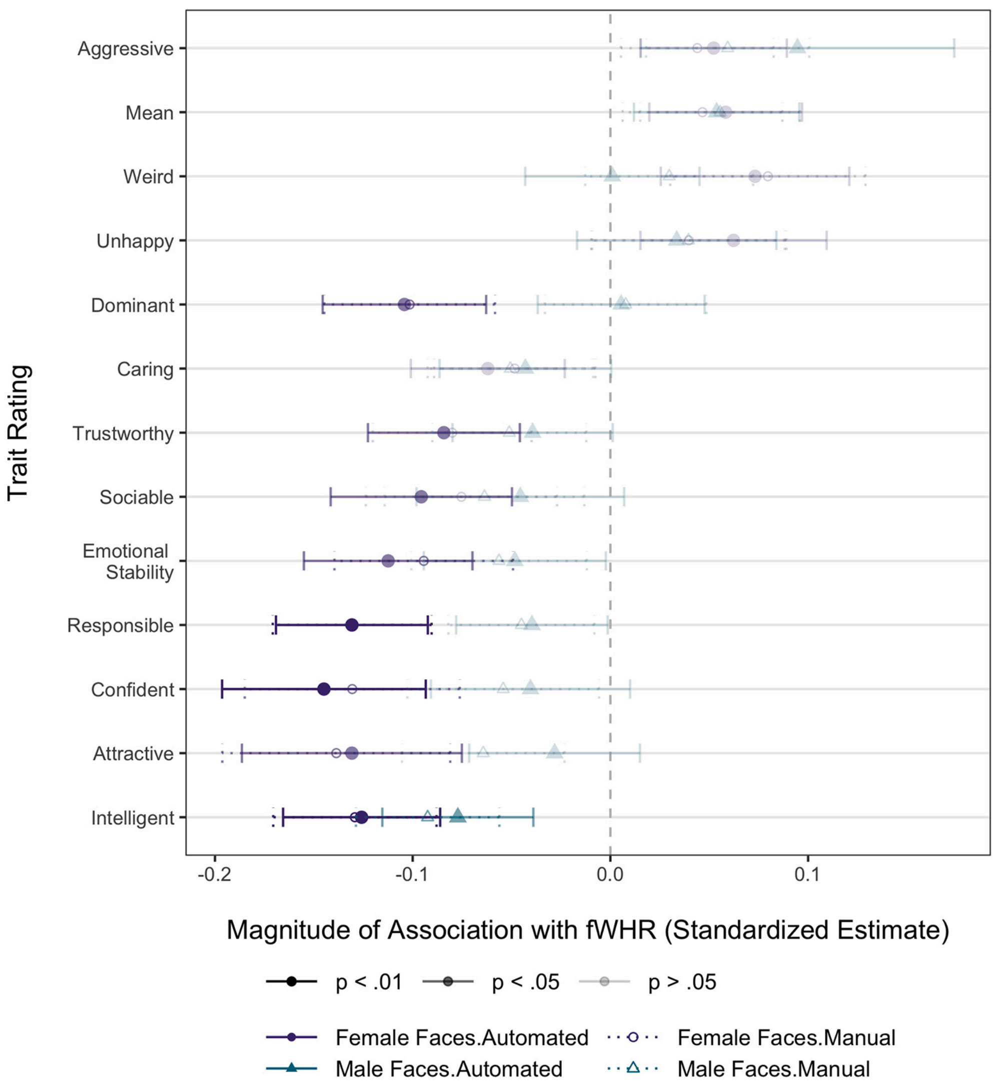


**Fig. 3.** Coefficient plot depicting the standardized estimates from multilevel models for the sex-specific associations between male ($n = 60$) and female ($n = 60$) facial width-to-height ratio (fWHR) and 13 trait inferences using fWHR scores based on manual measurement of faces (dotted lines and empty shapes) as well as a recently developed automated measurement method (solid lines and filled shapes). Error bars depict the standard error of the point estimates. The transparency of each point estimate and error bar depicts the level of statistical significance, where more transparency is associated with larger *p*-values and less transparency is associated with smaller *p*-values.

for each of the 120 faces would equal the number of raters who rated the trait in each country for the original study. This step essentially simulates a replication of the PSA rating study, assuming that the original ratings are representative of the population of raters.

2) we computed the average rating for each trait from these sampled ratings for each of the 120 faces; this step simulates what each face's average trait rating might be based on the new sample of raters.
3) these trait-scores were merged with the CFD norming data containing the averaged trait-scores for all 597 faces, and we again sampled with replacement from this merged dataset such that the number of male and female faces were equal to the original proportions. This step effectively simulates a new sample of targets, assuming that the CFD photoset represents the population.
4) we used the *lavaan* package (Rosseel, 2012) to set up structural equation models that simultaneously regressed the CFD trait ratings for the resampled full-set of faces and the PSA trait ratings for the resampled subset of faces onto the automated fWHR[5] estimates available for the full sample using FIML. This approach effectively zeroes in on what the association between fWHR and the PSA trait-ratings would need to be if full sample of CFD faces had been rated in the PSA study in order to reproduce the observed covariances between a) fWHR and the CFD trait ratings and b) the CFD trait ratings and the PSA trait ratings.
5) We iteratively saved the parameter estimates for the association between fWHR and trait ratings for each bootstrapped replicate to construct a distribution of the parameter estimates.

After the fWHR-inference associations were calculated for 5000 bootstrapped replicates for each sex and world region, we computed the mean and the upper 97.5% and lower 2.5% quantiles of the correlation-estimate distribution for each sex and trait combination within each world region. The resulting population correlation estimates and 95%

[5] We used the fWHR scores generated Jones et al. (2020) automated procedure because we did not have the resources to manually measure fWHR for all 597 CFD faces. Our previous analyses showed that the automated measure was strongly correlated with manual measurements on the subset of 120 faces, and that the fWHR-trait associations obtained using fWHR scores from the automated measurement procedure were not qualitatively different from those using the manual method in the subset of 120 faces.

confidence intervals for the FIML-estimated associations between fWHR and the conceptually-overlapping PSA and CFD trait-ratings are shown in Fig. 4.

In this larger set of faces, we obtained a somewhat different pattern of results than in our planned analyses with only 120 faces. Specifically, fWHR was reliably positively associated with PSA ratings of how mean and aggressive male faces but not female faces were perceived to be across most world regions. Dominance perceptions were negatively associated with fWHR in female faces, but the association was essentially null in male faces. Attractiveness perceptions were negatively associated with fWHR in female faces and male faces to about the same degree, but this association exhibited more variability across bootstrapped replicates for male faces. Finally, fWHR was reliably negatively associated with trustworthiness judgments of male faces, but this association was smaller and not reliably greater than zero across replicates for female faces. There does appear to be some regional variation in the associations between fWHR and trait judgments, but the variability is primarily in the relative magnitude of the effects rather than the direction of associations.

We also explored whether the fWHR-trait inference associations differed by ethnicity of target faces. We followed the same bootstrapping procedure as outlined above for the FIML-estimated associations between traits and fWHR, but we computed the association separately by ethnicity of the target and pooled ratings across world regions rather than computing them separately. The results of these analyses are shown in Fig. 5. The results tentatively suggest that target ethnicity may moderate the strength of the associations between fWHR and some trait inferences; (e.g., fWHR was more reliably associated with dominance inferences in White male faces than faces of other ethnicities). But most most of these differences were generally relatively small in magnitude, not consistently in the same direction, and not found reliably in both the PSA and CFD rating samples.

## 4. Discussion

We leveraged trait ratings of relatively diverse faces by large samples of raters from around the world to examine whether fWHR is reliably associated with social inferences that are foundational to social perception (Jones et al., 2021; Oosterhof & Todorov, 2008). To do so, we measured the fWHR of faces using a recently developed automated method (Jones et al., 2020) and conducted preregistered confirmatory analyses to examine relationships between fWHR and trait judgments, as well as exploratory analyses to probe the sensitivity of our results to the subset of faces that happened to be included in the cross-national study. Our results revealed mixed evidence for theoretically expected associations between fWHR and trait ratings—namely, that fWHR should be positively related to inferences of threat and dominance, and negatively related to inferences of trustworthiness and attractiveness in men (c.f., Geniole et al., 2015; Stirrat & Perrett, 2010)

In our preregistered multilevel models, we found that fWHR was significantly negatively associated with perceptions of female faces' dominance, trustworthiness, sociability, emotional stability, responsibleness, confidence, attractiveness, and intelligence. In contrast, fWHR of male faces was only weakly associated—if at all—with any trait inferences. Thus, these analyses provide little support for theoretical predictions. However, it is possible that the relatively small subset of faces used in these analyses do not provide adequate power to reach statistical significance for very small effects, and the subset of faces may not be representative of the larger set of faces from which they were drawn.

We found a somewhat different pattern of results in our exploratory analyses, which ameliorated power and representativeness issues by incorporating information from independent ratings of the larger sample of CFD faces and a bootstrapping procedure to factor in variability in PSA ratings from different world regions. Most pertinent to probing the role of fWHR in threat perception, fWHR was reliably positively associated with perceptions of meanness and aggressiveness in male faces,

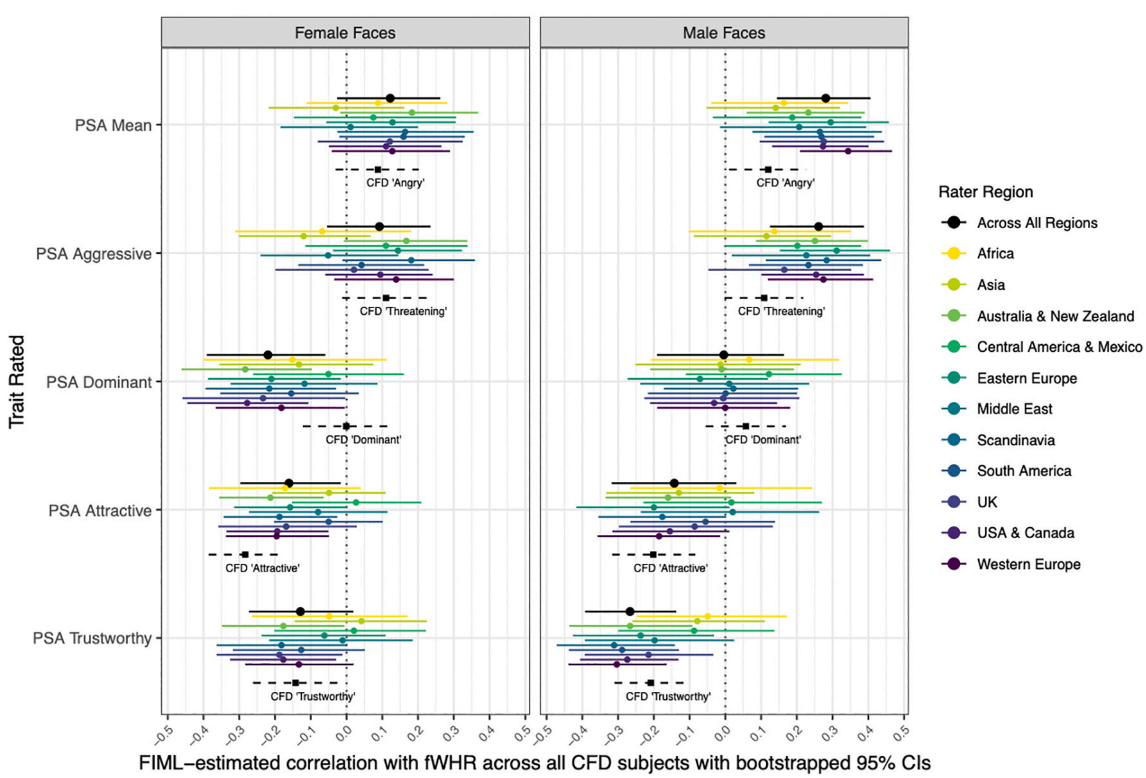


**Fig. 4.** Coefficient plot depicting the bootstrapped mean correlation estimate and 95% confidence intervals for the correlations between facial width-to-height ratio (fWHR) and aggregated ratings of traits inferred from male and female faces by raters across world regions.

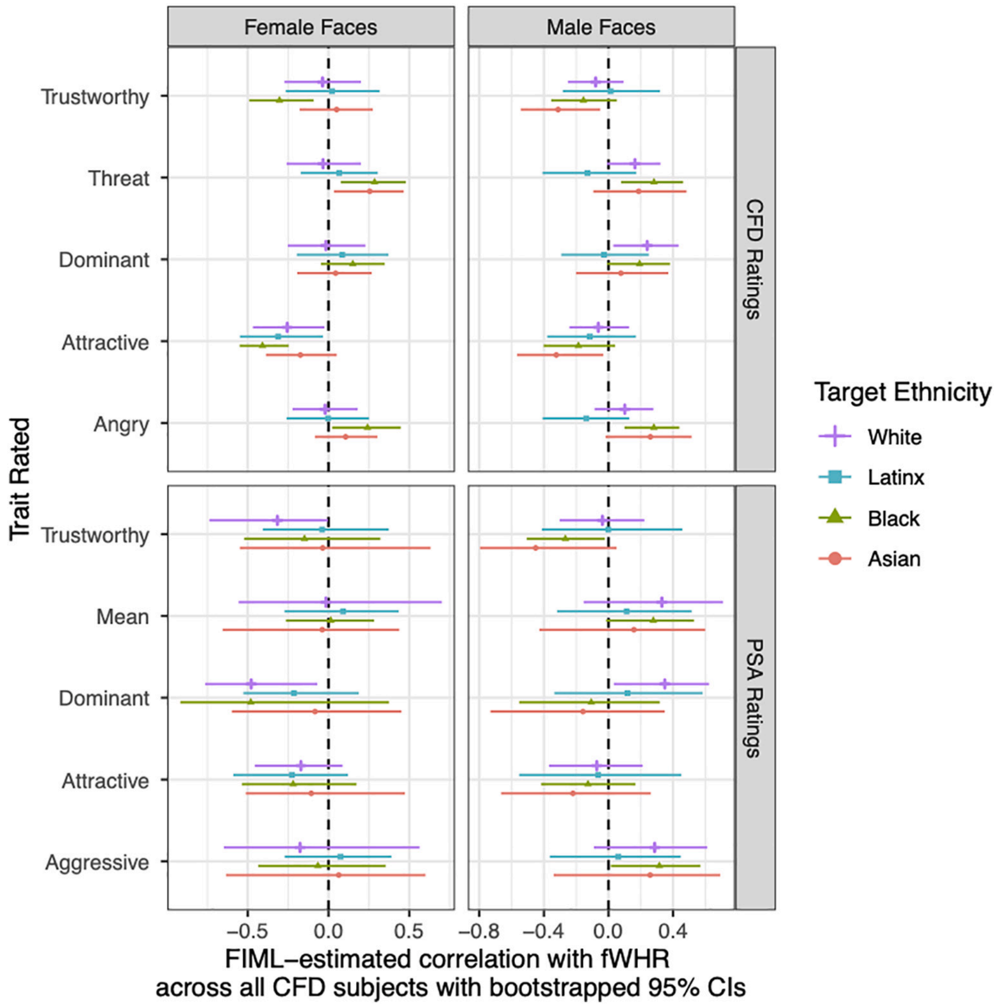


**Fig. 5.** Coefficient plot depicting the bootstrapped mean correlation estimate and 95% confidence intervals for the correlations between facial width-to-height ratio (fWHR) and aggregated ratings of traits inferred from male and female faces by raters in the original CFD norming dataset and the larger samples of PSA raters across world regions broken apart by the ethnicity of the faces being rated.

but less so in female faces. Additionally, fWHR was reliably negatively associated with dominance in female but not male faces, with trustworthiness in male faces but less so in female faces, and with attractiveness to about the same degree in both male and female faces (albeit less reliably so in male faces).

The finding that higher male fWHR is associated with perceptions of greater threat and lower trustworthiness in the full sample of faces is consistent with hypotheses that fWHR may be a cue to threat. However, mounting evidence suggests that there are no reliable links between fWHR, testosterone, and behavioral tendencies (e.g., Kordsmeyer et al., 2019; Kosinski, 2017; Wang et al., 2019). Thus, it is not clear why fWHR is associated with these dispositional inferences.

As we discussed in the introduction, fWHR-inference associations in the absence of commensurate fWHR-behavior associations may be a product of evolutionary mismatch or a byproduct of other functional mechanisms. To support the mismatch explanation, there is some comparative evidence in non-human primates to suggest that fWHR may be phylogenetically ancient cue of aggression and dominance (for review, see Wilson et al., 2020), but these associations run contrary to the null (male faces) and negative (female faces) associations between fWHR and dominance judgments in the current study. Given that previous research suggests that faces with higher fWHR are more likely to be perceived as angry (e.g., Deska et al., 2017; Neth & Martinez, 2010), it may be more plausible that fWHR-threat perceptions are a byproduct of anger-detection mechanisms. Specifically, individuals expressing neutral expressions with higher fWHR may be perceived as more likely to be angry, and thus more likely to be aggressive (Sell, 2011; Wyckoff, 2016) and potentially less trustworthy (Oosterhof & Todorov, 2009; Todorov, 2008)—especially because the zero-acquaintance paradigms typically employed in person-perception research lack the rich contextual information that would typically accompany social interactions. On this view, fWHR-based judgments could be the result of mechanisms that infer probable behaviors based on perceptions of emotions and their likely outputs in a context-specific manner (e.g., if a person looks angry, they may behave aggressively or defect in cooperative exchange), rather than mechanisms that use evolved priors about ancestrally-recurrent associations between static cues of developmental testosterone and general behavioral dispositions, which is the prominent theoretical explanation (e.g., Carré & McCormick, 2008). We are unable to test between byproduct and mismatch explanations with the current data, but future research on fWHR-inferences should consider teasing them apart when investigating links between fWHR and threat perceptions—keeping in mind that they are not necessarily mutually exclusive.

Our findings that fWHR was reliably negatively associated with dominance perceptions of female, but not male, faces may suggest that fWHR plays a larger role in social judgments of female faces than hypotheses rooted in the developmental effects of testosterone on fWHR would predict (e.g., Carré & McCormick, 2008). The few extant studies examining links between fWHR and behavioral tendencies in women have found mixed evidence of reliable fWHR-behavior links (e.g., Haselhuhn & Wong, 2011; Lefevre, Etchells, Howell, Clark, & Penton-Voak, 2014; Stirrat & Perrett, 2010). Large-scale research may be needed to accurately examine whether fWHR-based judgments track real relationships between fWHR and behavior in women. It may also be worthwhile to rigorously investigate hormonal correlates of fWHR in women, as this is an area which has been relatively understudied in comparison to men. Of course, it is possible that fWHR is not associated with behavioral or hormonal variation in women, and the fWHR-trait judgments also may be byproducts stemming, for example, from face-based assessments of age, sex-typicality, body size, or other assessments that might influence attractiveness perceptions (Coetzee, Chen, Perrett, & Stephen, 2010; Fiala et al., 2020).

Our exploratory analyses are suggestive that face-ethnicity may moderate relationships between trait inferences and fWHR, but these interaction effects are likely to be small. Given that the sample sizes of each ethnicity in the full face database are still relatively small and unbalanced when broken apart by sex (*n*s range from 52 to 108), these differences could largely reflect sampling variability rather than true effects and should be interpreted with caution. In our view, there is little reason to expect selection to drive differential associations between fWHR and trait inferences by ethnicity or race per se (Kurzban, Tooby, & Cosmides, 2001; Pietraszewski, 2021), but manifest differences may be reflective of biased emotion detection accuracy in faces that are perceived to be potential outgroup members (e.g., Becker, Neel, & Anderson, 2010; Halberstadt et al., 2020; Hugenberg & Bodenhausen, 2003; Miller, Maner, & Becker, 2010). Still, our estimates of the small differences in fWHR-inferences associations across ethnicities can inform future research investigating the extent to which face-ethnicity moderates fWHR-inferences, suggesting that large samples of diverse faces will be required to reliably detect the likely small interactions.

More generally, our findings of small, inconsistent effects could be interpreted as evidence that fWHR is not major cue used in person perception. Evolutionary mismatch and byproduct explanations may be unnecessary, unless small associations between fWHR and trait inferences are meaningful in social cognition. To better understand the magnitude of the relationships between fWHR and trait inferences, future research should consider how small fWHR-inference associations can be while still being considered theoretically relevant or interesting, and aim to use larger samples of faces in order to be maximally powered to detect effects at least as small as those found in the current study. Adequate power can be efficiently achieved for future studies by using a large number of face stimuli in a stimuli-within-block design where different groups of raters judge different sets of photos (Westfall et al., 2014).

Several limitations should be considered in interpreting our findings, in addition to power and stimuli-representativeness that we attempted to address and discuss previously. Some research suggests that body size or BMI may confound fWHR associations (e.g., Deaner, Goetz, Shattuck, & Schnotala, 2012; Geniole et al., 2015), but we could not examine interactions or control for allometric scaling because no data on body size were available for the faces used as stimuli. Relatedly, no behavioral data were associated with targets, so we could not assess the accuracy of rater judgments. Additionally, minimal data are available on rater characteristics, so we could not examine interactions between potentially relevant rater characteristics and judgments of faces, which may have moderating effects (e.g., Gruenewald, Kemeny, & Aziz, 2006; Tiedens, Unzueta, & Young, 2007). Finally, although representative of many nations and cultures, the raters in the sample are still mostly young college students, so our findings may not generalize well to older adults or to small-scale, nonindustrial societies.

## 5. Conclusion

Identifying the features correlated with face-based social judgments may help us better understand the mechanisms behind person perception. Although fWHR initially appeared to be a promising facial feature for studying person perception, the inconsistent and contradictory evidence indicates that the underlying theoretical framework needs modification. Our findings suggest that resources could be fruitfully directed towards rigorous tests of fWHR-links in women, examination of other morphological features involved in face perception, and investigation into the degree to which robust associations between fWHR and trait inferences reflect evolutionary mismatch or byproducts.

## Authors' contributions

Both authors were involved in conceptualization. PD conducted formal analyses and made visualizations. Both authors contributed to the writing the original draft, as well as reviewing and editing.

## Acknowledgements

We are grateful to the Psychological Science Accelerator for curating the rating data, all the member labs that were involved in collecting the data we used in this examination, and Cierra Flood, Danielle Patterson, and Dylan Stephenson from the ASU Cooperation and Conflict lab for their assistance in manual fWHR measurement. We also thank David M. Buss and Courtney L. Crosby for providing helpful feedback on an earlier draft of this paper.

## References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Becker, D. V., Neel, R., & Anderson, U. (2010). Illusory conjunctions of angry facial expressions follow intergroup bias. *Psychological Science, 21*(7), 938–940.

Bird, B. M., Jofré, V. S. C., Geniole, S. N., Welker, K. M., Zilioli, S., Maestripieri, D., … Carre, 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.

Boshyan, J., Zebrowitz, L. A., Franklin, R. G., Jr., McCormick, C. M., & Carré, J. M. (2014). Age similarities in recognizing threat from faces and diagnostic cues. *Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 69*(5), 710–718.

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. *Proceedings of the Royal Society B: Biological Sciences, 275*(1651), 2651–2656.

Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science, 20*(10), 1194–1198.

Coetzee, V., Chen, J., Perrett, D. I., & Stephen, I. D. (2010). Deciphering faces: Quantifiable visual cues to weight. *Perception, 39*(1), 51–61.

Deaner, R. O., Goetz, S. M., Shattuck, K., & Schnotala, T. (2012). Body weight, not facial width-to-height ratio, predicts aggression in pro hockey players. *Journal of Research in Personality, 46*(2), 235–238.

Deska, J. C., Almaraz, S. M., & Hugenberg, K. (2017). Of mannequins and men: Ascriptions of min in faces are bounded by perceptual and processing similarities to human faces. *Social Psychological and Personality Science, 8*(2), 183–190.

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.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*(1), 128–141.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430–457.

Fiala, V., Třebický, V., Leongómez, J., Tureček, P., Pazhoohi, F., Saribay, S. A., … Kleisner, K. (2020, April 16). *Cues for facial attractiveness and preference of sexual dimorphism: A comparison across five cultures*. https://doi.org/10.31234/osf.io/vh4rj.

Frankenshuis, & Del Giudice. (2012). When do adaptive developmental mechanisms yield maladaptive outcomes? *Developmental Psychology, 48*(3), 628–642.

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), Article e0132726.

Geniole, S. N., Keyes, A. E., Carré, J. M., & McCormick, C. M. (2014). Fearless dominance mediates the relationship between facial width-to-height ratio and willingness to cheat. *Personality and Individual Differences, 57*, 59–64.

Geniole, S. N., & McCormick, C. M. (2015). Facing our ancestors: Judgements of aggression are consistent and related to the facial width-to-height ratio in men irrespective of beards. *Evolution and Human Behavior, 36*(4), 279–285.

Gruenewald, T. L., Kemeny, M. E., & Aziz, N. (2006). Subjective social status moderates cortisol responses to social threat. *Brain, Behavior, and Immunology, 20*(4), 410–419. https://doi.org/10.1016/j.bbi.2005.11.005.

Halberstadt, A. G., Cooke, A. N., Garner, P. W., Hughes, S. A., Oertwig, D., & Neupert, S. D. (2020). Racialized emotion recognition accuracy and anger bias of children's faces. *Emotion*, 1–15. https://doi.org/10.1037/emo0000756.

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), Article e0122637.

Haselhuhn, M. P., & Wong, E. M. (2011). Bad to the bone: facial structure predicts unethical behaviour. *Proceedings of the Royal Society of London B: Biological Sciences, 279*, 571–576.

Haselhuhn, M. P., Wong, E. M., Ormiston, M. E., Inesi, M. E., & Galinsky, A. D. (2014). Negotiating face-to-face: Men's facial structure predicts negotiation performance. *The Leadership Quarterly, 25*(5), 835–845.

Hehman, E., Leitner, J. B., & Freeman, J. B. (2014). The face-time continuum: Lifespan changes in facial width-to-height ratio impact aging-associated perceptions. *Personality and Social Psychology Bulletin, 40*, 1624–1636.

Hodges-Simeon, C. R., Sobraske, K. N. H., Samore, T., Gurven, M., & Gaulin, S. J. (2016). Facial width-to-height ratio (fWHR) is not associated with adolescent testosterone levels. *PLoS One, 11*(4), Article e0153083.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science, 14*(6), 640–643.

Jones, A. L., Schild, C., & Jones, B. C. (2020). Facial metrics generated from manually and automatically placed image landmarks are highly correlated. *Evolution and Human Behavior, 42*(3), 186–193.

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., … Sirota, M. (2021). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour, 5*(1), 159–169.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54–69.

Köllner, M. G., Janson, K. T., & Schultheiss, O. C. (2018). Commentary: Sexual dimorphism of facial width-to-height ratio in human skulls and faces: A meta-analytical approach. *Frontiers in Endocrinology, 9*. https://doi.org/10.3389/fendo.2018.00227.

Kordsmeyer, T. L., Freund, D., Pita, S. R., Jünger, 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.

Kosinski, M. (2017). Facial width-to-height ratio does not predict self-reported behavioral tendencies. *Psychological Science, 28*(11), 1675–1682.

Kramer, R. S. (2015). Facial width-to-height ratio in a large sample of commonwealth games athletes. *Evolutionary Psychology, 13*(1), 197–209.

Kramer, R. 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.

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences, 98*(26), 15387–15392.

Kvarven, A., Strømland, E., & Johannesson, M. (2019, March 9). *Comparing meta-analyses and pre-registered multiple labs replication projects*. https://doi.org/10.31219/osf.io/brzwt.

Lassek, W. D., & Gaulin, S. J. (2019). Evidence supporting nubility and reproductive value as the key to human female physical attractiveness. *Evolution and Human Behavior, 40*(5), 408–419.

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.

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.

Lefevre, C. E., Lewis, G. J., Bates, T. C., Dzhelyova, M., Coetzee, V., Deary, I. J., & Perrett, D. I. (2012). No evidence for sexual dimorphism of facial width-to-height ratio in four large adult samples. *Evolution and Human Behavior, 33*(6), 623–627.

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.

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.

Li, N. P., van Vugt, M., & Colarelli, S. M. (2018). The evolutionary mismatch hypothesis: Implications for psychological science. *Current Directions in Psychological Science, 27*(1), 38–44.

Lick, D. J., & Johnson, K. L. (2018). Facial cues to race and gender interactively guide age judgments. *Social Cognition, 36*, 497–516.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*(4), 1122–1135.

Mason, M. F., Cloutier, J., & Macrae, C. N. (2006). On construing others: Category and stereotype activation from facial cues. *Social Cognition, 24*(5), 540–562.

Mileva, V. R., Cowan, M. L., Cobey, K. D., Knowles, K. K., & Little, A. C. (2014). In the face of dominance: Self-perceived and other-perceived dominance are positively associated with facial-width-to-height ratio in men. *Personality and Individual Differences, 69*, 115–118.

Miller, S., Maner, J., & Becker, D. V. (2010). Self-protection biases in group categorization: Threat cues shape the psychological boundaries between "us" and "them". *Journal of Personality and Social Psychology, 99*(1), 62–77.

Neth, D., & Martinez, A. M. (2010). A computational shape-based model of anger and sadness justifies a configural representation of faces. *Vision Research, 50*(17), 1693–1711.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087–11092.

Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion, 9*(1), 128.

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.

Özener, B. (2012). Facial width-to-height ratio in a Turkish population is not sexually dimorphic and is unrelated to aggressive behavior. *Evolution and Human Behavior, 33*(3), 169–173.

Pietraszewski, D. (2021). The correct way to test the hypothesis that racial categorization is a byproduct of an evolved alliance-tracking capacity. *Scientific Reports, 11*(1), 1–11.

Pinker, S. (2012). *The better angels of our nature: Why violence has declined*. Penguin Group USA.

Robertson, J. M., Kingsley, B. E., & Ford, G. C. (2017). Sexually dimorphic faciometrics in humans from early adulthood to late middle age: Dynamic, declining, and differentiated. *Evolutionary Psychology, 15*(3), 1–10.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Sell, A. N. (2011). The recalibrational theory and violent anger. *Aggression and Violent Behavior, 16*(5), 381–389.

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

Stirrat, M., Stulp, G., & Pollet, T. V. (2012). Male facial width is associated with death by contact violence: Narrow-faced males are more likely to die from contact violence. *Evolution and Human Behavior, 33*(5), 551–556.

Tiedens, L. Z., Unzueta, M. M., & Young, M. J. (2007). An unconscious desire for hierarchy? The motivated perception of dominance complementarity in task partners. *Journal of Personality and Social Psychology, 93*, 402–414.

Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences, 1124*(1), 208–224.

Todorov, A., Mende-Siedlecki, P., & Dotsch, R. (2013). Social judgments from faces. *Current Opinion in Neurobiology, 23*(3), 373–380.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*, 519–545.

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

Wang, D., Nair, K., Kouchaki, M., Zajac, E. J., & Zhao, X. (2019). A case of evolutionary mismatch? Why facial width-to-height ratio may not predict behavioral tendencies. *Psychological Science, 30*(7), 1074–1081.

Weston, E. M., Friday, A. E., & Liò, P. (2007). Biometric evidence that sexual selection has shaped the hominin face. *PLoS One, 2*(8).

Wilson, V., Weiss, A., Lefevre, C. E., Ochiai, T., Matsuzawa, T., Inoue-Murayama, M., … Altschul, D. (2020). Facial width-to-height ratio in chimpanzees: Links to age, sex and personality. *Evolution and Human Behavior, 41*(3), 226–234.

Wyckoff, J. P. (2016). Aggression and emotion: Anger, not general negative affect, predicts desire to aggress. *Personality and Individual Differences, 101*, 220–226.