Sexual dimorphism of facial width-to-height ratio in human skulls and faces: A meta-analytical approach

Robin S. S. Kramer*

Department of Psychology, University of York, York, YO10 5DD, UK

* Corresponding author.

E-mail address: remarknibor@gmail.com (R. S. S. Kramer).

Word count: 5891

Acknowledgements

I am very grateful to Lyman Jellema for providing access to the Hamann-Todd human osteological collection, as well as background information and other details regarding the specimens. I also thank Rolando González-José for providing the data that he and his colleagues had previously collected. Finally, I thank Geoff Cumming for statistical advice, and Alex Jones and Renée Lefebvre for comments on the manuscript.

Abstract

Facial width-to-height ratio (FWHR), defined as the width of the face divided by the upper facial height, is a cue to behaviour. Explanations for this link often involve the idea that FWHR is sexually dimorphic, resulting from intersexual selection pressures. However, few studies have considered sexual dimorphism in skulls since the original paper on this topic, and it is possible that different explanations may be required if faces show sex differences but skulls do not. Here, meta-analyses of skulls found that men did have larger FWHR than women, although this effect was small. However, after categorising samples by ethnicity and geographical origin, meta-analyses only found evidence of sex differences in East Asians, and again, this effect was small. A re-analysis of previous studies after excluding skull samples found little evidence of sexual dimorphism in faces. Again, considering ethnicities separately, I found no differences for White samples but a medium-sized effect with East Asians, although this was not statistically significant with only three samples. Taken together, I found no reason to consider FWHR as a sexually dimorphic measure in skulls or faces, at least not universally, and so accounts based upon this assumption need rethinking if researchers are to explain the relationship between FWHR and behaviour.

Keywords: Facial width-to-height ratio; Sexual dimorphism; Skulls; Faces; Testosterone

1. Introduction

The idea that the human face provides social information is not a new one (Darwin, 1872). We can determine the identity (Bruce & Young, 1986), sex (Burton et al., 1993), age

(Rhodes, 2009), and ethnicity (Montepare & Opeyo, 2002) of a stranger with relative ease, as well as more dynamic and changing information like emotional state (Elfenbein & Ambady, 2002). There is also evidence that trait information like personality, physical and mental health, and even sexual orientation can be perceived with some accuracy from faces alone (Jones et al., 2012; Kramer & Ward, 2010; Rule et al., 2009; Scott et al., 2013).

In 2007, researchers provided evidence of one particular facial measure, the width-to-height ratio (FWHR – see Fig.1; Weston et al., 2007), which they found to be sexually dimorphic in human skulls, and has since been the subject of intense investigation as a cue to numerous behaviours. While overall size differences play a large role in general skull dimorphism (Calcagno, 1981; Lestrel, 1974; Rightmire, 1970), Weston and colleagues suggested that this ratio difference was instead due to developmental differences in shape trajectories during puberty. Specifically, the height of the upper face (defined as the nasion-prosthion distance) in adults is similar in men and women, while the (bizygomatic) width is larger in men. In other words, while sex differences in skulls are expected simply because men grow to be larger than women, the bizygomatic width in males shows additional growth at puberty beyond this predicted increase. The researchers argued that this difference in skull shape might result from intersexual selection pressures, so that a region of the face has evolved which highlights the distinction between men and women.

Why evidence of sexual dimorphism predicts an association between FWHR and behaviour is less clear. If female preferences led to increased facial width in men (although evidence actually suggests that wider faces are judged to be less attractive; Geniole et al., 2015), it may not necessarily follow that within-sex differences are correlated with behaviours. More intuitively, intrasexual selection pressures (e.g., male-male competition) could have resulted in increased success for wider-faced men, resulting in an appearance–behaviour link, especially if these two factors have the same underlying mechanism (testosterone, for instance). Might both explanations overlap, whereby a facial cue that highlights ‘maleness’ to women has become associated with masculinity (both in appearance and behaviour) in men? Of course, there is no reason to assume that the mechanisms underlying sex differences in facial development are the same as those that may drive a within-sex association between appearance and behaviour.

While the precise account and its relationship with sexual dimorphism remains unclear, FWHR does appear to function as a social cue. Levels of masculine characteristics (e.g., aggression, dominance, deception) in men correlate with FWHR, as do perceived levels of these traits (for meta-analyses, see Geniole et al., 2015; Haselhuhn et al., 2015). The explanation for this FWHR–behaviour association is thought to involve testosterone (Carré & McCormick, 2008; Sell et al., 2009), which may influence both facial development and behavioural characteristics. Indeed, initial research found significant associations between FWHR in men and baseline levels of testosterone, as well as testosterone changes in response to potential mate exposure (Lefevre et al., 2013).

Somewhat problematically for this account, FWHR may not actually be sexually dimorphic in faces (Kramer et al., 2012; Lefevre et al., 2012; Özener, 2012) or skulls (Gómez-Valdés et al., 2013; Stirrat et al., 2012). Of course, it may be that different mechanisms drive facial development in men and women, allowing for testosterone-produced correlates of behaviour in men without differences between the sexes (Lefevre et al., 2013). In a recent meta-analysis of this field, the authors found significant (but small) sex differences when considering studies of both skulls and faces together (Geniole et al., 2015), as well as for subsets of studies (2D photographs versus other materials). However, it is not clear whether differences remain when only skulls are analysed since this distinction was not made in their analyses. It may be that skulls do not show sex differences in FWHR but faces do, perhaps through evolved cues that utilise soft tissue deposits, which differ in men and women (Enlow, 1982). This would be an important caveat when investigating the explanatory mechanisms linking behaviour and facial measures.

One potential issue with previous investigations is that they have not considered populations separately based upon ethnicity or geographical origin. Given evidence of between-population differences in skulls (Gill & Rhine, 2004; İşcan & Steyn, 1999; Ousley et al., 2009), the inclusion of all groups into a single analysis will inherently suffer from this additional source of noise. It may be that FWHR dimorphism is present in some ethnicities/populations but not others, and this could account for the mixed results that have previously been found with faces. This would also be an important caveat for theories of dimorphism and signalling.

The other problem for the 'FWHR–testosterone–behaviour' account is that FWHR may not actually be associated with testosterone. In recent research investigating several samples and reporting a combined meta-analysis, no relationship was found between FWHR in adult men and baseline testosterone or competition-induced testosterone reactivity (Bird et al., 2016). Even during adolescence, when testosterone is hypothesised to impact facial growth (Weston et al., 2007), no relationship was found between male FWHR and testosterone levels or other known testosterone-derived traits (Hodges-Simeon et al., 2016). Indeed, FWHR showed no change during adolescence and no growth spurt, contrary to predictions.

In the current work, I focussed specifically on whether FWHR is sexually dimorphic in adult human skulls using a meta-analytical approach. Given that the popular topic of FWHR as an important facial cue originated from this initial finding (Weston et al., 2007), it is worth further examination using multiple large samples. I also considered geographical and ethnic origins as potential factors in order to allow for the likelihood that populations may differ. For this reason, I revisit the topic of FWHR sex differences in faces, again considering ethnicity as a potential influence. Importantly, prior large-scale analyses in this area have yet to consider the distinction between faces and skulls, and the possibility (and evolutionary implications) that there may be FWHR sex differences in one but not the other.

2. Methods

2.1. Previous research

All peer-reviewed and published manuscripts that investigated human skull FWHR separately for men and women were included. This involved searching through all articles that cited Weston et al. (2007), the first paper to propose this measure as a topic of interest. Conveniently, all articles prior to the end of 2014 had already been identified in the recent meta-analysis by Geniole et al. (2015), and no newer research (as of May 2016) or omissions were found. This resulted in the inclusion of three peer-reviewed manuscripts.

In total, these publications described eight separate skull databases: six (Gómez-Valdés et al., 2013), one (Stirrat et al., 2012), and one (Weston et al., 2007). Problematically, the authors reported, and the previous meta-analysis utilised, summary database values for FWHR rather than separating these into specific populations in terms of origin/ethnicity. For example, Stirrat and colleagues provided means and standard deviations for their full sample, which included a mixture of White and non-White skulls. Similarly, Gómez-Valdés and colleagues reported the average FWHR dimorphism for each database, which did not allow for the analysis of separate populations, incorporating different sample sizes, etc. For instance, the Howells (1973, 1989, 1995) database alone contained 26 groups (of varying sizes) originating from almost as many countries.

To address this issue, I contacted the authors (Gómez-Valdés et al., 2013) and obtained summary statistics for their databases, separately for each population. This would allow the calculation of an effect size for each group rather than each database. Unfortunately, the authors were unable to provide data regarding two of their previously reported databases (Hallstat and Mexico City Penitentiary) due to ethical constraints and data property issues, and so these two sets were not included in the current meta-analysis. In addition, several populations were removed before analyses because there was substantial overlap across their databases. For example, the Ourga specimens appeared in both the 2D and Pucciarelli sets. Whenever multiple occurrences were found, the repeated case with the smaller sample size was removed. This was because the second appearance often featured fewer specimens and so was assumed to be a subset of the larger sample, and this was confirmed by the authors through correspondence.

For Weston's original sample (Weston et al., 2007), the set contained individuals from several different southern African populations. However, each of these was not represented in sufficient numbers to allow separation into subgroups, and the authors reported previous work demonstrating that these populations were comparable (de Villiers, 1968). I therefore considered this set as a single population for the purposes of analysis.

Finally, in order to allow for populations as separate sets/studies, I incorporated into the meta-analysis only groups that included a minimum of two men and two women. Additional data/populations were discarded.

2.2. Databases

Although Gómez-Valdés et al. (2013) provided their summary data regarding the Howells database (Howells, 1973, 1989, 1995), I was able to obtain this set independently (see Section 2.2.1). I therefore used my own calculated values for these populations, given that it was preferable to work with the raw data when available. Similarly for the database used by Stirrat et al. (2012), I obtained the original set independently (see Section 2.2.2). From these data, I was able to calculate summary statistics separately for each population.

In addition, I obtained four new skull databases in order to increase the number of populations included in the meta-analysis and improve the reliability of the findings. A full summary of the final databases included in the analysis can be found in the Supplementary Materials.

2.2.1. William W. Howells craniometric data set

This database contained information on a large number of specimens. All skulls were from adults (approximately 18 years old and above, as determined by dental development, although exact age was not known), and sex and origin were included. The skulls were pooled from historical collections from various institutions internationally, and contained 30 indigenous populations. A full description can be found in Howells' monographs (1973, 1989, 1995). FWHR was calculated using the bizygomatic breadth and nasion-prosthion height, measured directly from the skulls. The usable data included here comprised 2412 individuals from 26 populations.

2.2.2. Database for forensic anthropology in the United States, 1962-1991

This forensic database was created in order to represent the ethnic diversity and demographic structure of the US population. A full description can be found in the codebook (Jantz & Moore-Jansen, 2006). From the initial set, specimens were excluded due to missing cranial measures, if they were aged below 18, or if their sex or race were not reported. Given that the current analysis relies heavily upon the accuracy of these two pieces of information, I also chose to exclude specimens where there was a label for sex/race but the researchers had specified uncertainty in their reporting of these categories. The usable data included here comprised 665 individuals from two populations.

2.2.3. Hamann-Todd human osteological collection

This database contained skulls collected in Cleveland, Ohio, and the surrounding area in the first half of the twentieth century, and housed at the Cleveland Museum of Natural History. The sex, age, and ethnicities of cadavers were recorded, along with skull measurements. Those specimens where age, sex, or the necessary measures were missing, were excluded from analyses, along with any specimens younger than 18 years old. As above, FWHR was calculated using the bizygomatic breadth and nasion-prosthion height, measured directly from the skulls. The usable data included here comprised 2614 individuals from three populations.

2.2.4. Forensic 3D database

A database of specimens was created in order to facilitate forensic identification using geometric morphometrics. The set incorporates skulls taken from several different collections, including the Roger J. Terry anatomical skeletal collection (Smithsonian Institution, Washington, DC) and the forensic data bank (University of Tennessee, Knoxville), and features populations from around the world. All specimens were adults over the age of 18 (determined based on standard growth and development). Sex and origin information was available, along with three-dimensional coordinates for craniofacial landmarks. These were used to calculate bizygomatic breadth and nasion-prosthion height. The usable data included here comprised 419 individuals from six populations.

2.2.5. "Hispanic" populations craniometric database

This database was created by the North Carolina State University's forensic analysis lab in order to investigate sources of admixture in "Hispanic" populations, and represents individuals from European, South and Central American countries. Sex, age, and information regarding skull width and height, were missing for many of the items. In addition, specimens under the age of 18 (determined based on standard growth and development) were also excluded. Finally, all specimens originating from Peru were removed since a subset of these also appeared in the Howells database. The usable data included here comprised 50 individuals from two populations.

2.2.6. Modern, cranial, postcranial and dental metrics database

The majority of these measurements were recorded by Peter Brown at Australian National University. The sample consists of Australian Aborigines, Southern Chinese, and European skulls. Many of the specimens have missing data, including age (although all were classed as adults). Bizygomatic breadth and nasion-prosthion height were measured directly from the skulls. The usable data included here comprised 259 individuals from five populations.

2.3. Ethnicity

Populations were broadly categorised where sufficient numbers were present. I make no claims about the nature of race as a useful, or even justifiable, biological concept (Smedley & Smedley, 2005). Instead, simply in order to investigate whether populations may show different patterns based on their similar ethnic or geographical origins, I used broad umbrella terms that may help to highlight commonalities and differences in skull sexual dimorphism. For example, I used 'White' to incorporate individuals originating from Europe and North America, where their origin was considered to be Caucasian. Similarly, 'Black' included both North American and African skulls that had been previously classified as Black, African American or Native African. I also used the broad categories 'East Asian', 'Australian Aboriginals', 'Pacific Islands', and 'South American'.

It is important to note that craniometric variation at the global level (between geographic regions or populations) is much lower than within local populations (Relethford, 2002). Indeed, perhaps as low as 11-14% of global diversity exists between regions, where the rest falls within regions (Relethford, 1994). Therefore, it would be preferable to categorise the current samples using far narrower groupings than the ones presented here, focussing on true populations (e.g., Han Chinese) rather than more general regions (East Asian), as this would likely improve the chances of finding group-level differences. Unfortunately, the availability of samples prevents such narrow categorisations while still maintaining a reasonable subgroup size. However, such within-subgroup noise means that any statistical differences between groups are perhaps even more suggestive of ethnic/population differences.

2.4. Statistics

All data were based upon differences between men and women, and so I chose to use Cohen's *d* as the effect size. Analyses were carried out using customised Microsoft Excel spreadsheets, based upon suggestions outlined in previous work (Geniole et al., 2015), as well as the formulae and guidelines provided by Cumming (2012). Specifically, the pooled estimate of the standard deviation within groups was used as the standardiser for *d*. In addition, unbiased estimates of δ (d_{unb}) were used in all cases after applying Hedges' adjustment to *d* to account for small samples [$d*(1-(3/((4*df) – 1)))$ where *df* is degrees of freedom]. Finally, the effect size for each dataset was weighted by the inverse of its variance before calculation of the mean weighted effect size.

The 95% confidence intervals for each study's effect size (see Supplementary Materials) were calculated (using Wilsons's online calculator: http://www.campbellcollaboration.org/escalc/) around *d* rather than d_{unb} because these provide a better estimate of the intervals around the population value, δ (Cumming, 2012).

All analyses presented here use random effects models, which assume that the population means estimated by the different studies are randomly chosen from a superpopulation (heterogeneity). Fixed effects models, in contrast, assume that every study estimates the same mean (homogeneity), and so any variation in sample effects is due to sampling error alone. Although random effects models are more complex, they are also considered more realistic and are generally recommended (Cumming, 2012). One further advantage is that fixed effects models are simply a special case of random effects models, where the population variance happens to be zero (Hunter & Schmidt, 2004). As discussed below, I also found statistical evidence to suggest that the samples were heterogeneous.

3. Results

3.1. Meta-analysis of skulls for all populations

The Supplementary Materials provides a summary of the eleven databases included in the meta-analysis, comprising 4918 men and 2924 women from 87 populations.

3.1.1. Heterogeneity

The two previous meta-analyses in this field disagreed with regard to which type of model was most suitable: fixed (Haselhuhn et al., 2015) or random effects (Geniole et al., 2015). As such, I first discuss the evidence supporting the use of random effects models here.

Several statistics were considered in order to explore the heterogeneity of the databases (whether different samples estimate different population effect sizes or a single one). First, I found statistically significant variability between study means, $Q(86) = 162.01$, $p < .0001$. In other words, the observed variation across studies (162) was greater than the expected variation (which is equal to the degrees of freedom, 86). However, this test can be both poor at detecting true heterogeneity due to low power, and can have excessive power with many/larger studies (Higgins et al., 2003). As such, other measures (e.g., T^2 or I^2) often prove more informative with regard to the amount of inconsistency, but can also allow comparisons to be made across analyses. The estimated variance of the true effect sizes (the amount of true heterogeneity) appears relatively low ($T^2 = 0.05$), although notice that this means our estimate of their *SD* is 0.22 while the mean effect size itself (see Section 3.1.2) is only 0.09 in the same units. Further, about half ($I^2 = 46.92\%$, considered moderately large) of this observed variance is real, i.e., due to heterogeneity rather than simply being spurious. Finally, the 'diamond ratio' (Cumming & Calin-Jageman, 2017), calculated by dividing the margin of error produced by the random effects model by the one given by the fixed effects model, was

1.59. Since this is a ratio, a value of 1 would suggest little heterogeneity, and so the current value implies there is heterogeneity present. Taken together, there is evidence here to proceed with random effects models, which indeed many recommend the use of in all situations (Cumming, 2012).

3.1.2. Results

The results of the meta-analysis found that men's FWHR was slightly larger than women's, $N = 7941$, $k = 87$, mean weighted $d = 0.09$, 95% CI [0.01, 0.17], $p = .022$. This result is in line with the previous findings of Geniole et al. (2015), whose effect size was 0.11, [0.03, 0.20], and included studies measuring both skulls and faces.

Inspection of the 87 samples (see S1 Fig) identified eight apparent outliers, which had effect sizes with confidence intervals that did not overlap with those of the mean weighted effect size. (These are noted in the Supplementary Materials.) Excluding these outlying effect sizes increased the mean weighted effect size and decreased the confidence interval, $N = 5955$, $k = 79$, mean weighted $d = 0.10$, [0.04, 0.16], $p = .002$. In addition, the variability between study means was no longer significant, $Q(78) = 96.93$, $p = .072$. Finally, the variance due to heterogeneity could now be considered small, $I^2 = 19.53\%$.

As noted above, 47% of the variation across samples was due to heterogeneity rather than chance (prior to the removal of outliers). Given this degree of variability, it may be the case that one or more study-level characteristics (moderators) could account for some of this variation. First, I consider ethnicity as a potential moderator.

3.2. Meta-analyses of skulls by ethnicity

3.2.1. Subgroup analysis

I carried out a subgroup analysis to investigate whether ethnicity was a moderator in the overall meta-analysis. Similar to a conventional analysis of variance, this method allows for the comparison of effect sizes across subgroups (here, ethnicities) in order to determine the effect of group-level variables. Study samples were labelled as one of six broad categories of ethnicity/origin, excluding the remaining populations that did not fall within one of these categories. A summary of these can be seen in Table 1.

For random effects models, I need to estimate the value of τ^2, the variance of true effect sizes across the set of studies/samples. Since I am interested in estimating the mean and sampling distribution for each subgroup, I need an estimate of τ^2 within each subgroup. There is no *a priori* reason to assume that the true study-to-study dispersion is the same within all subgroups, and so I use a separate estimate of τ^2 for each subgroup. However, if there are only a few studies within subgroups (e.g., fewer than five; see Table 1) then the estimates of τ^2 are likely to be imprecise. In such cases, the recommendation is to use a pooled estimate in order to increase accuracy (Borenstein et al., 2009). Here, I present the results of both methods.

Using random effects with separate estimates of τ^2, I found that ethnicity was not a statistically significant moderator, $Q_{\text{between}}(5) = 7.51$, $p = .186$. Utilising a pooled estimate of τ^2 produced a similar result, $Q_{\text{between}}(5) = 4.80$, $p = .440$. However, the ability to demonstrate that ethnicity is a moderator in these analyses requires large variation between subgroup means and little variation within subgroups. Problematically, at least some of the subgroups show large within-group variation (see Table 1), making any moderator effects difficult to detect.

One way to address this large within-subgroup variation is to remove any outlying effect sizes. As with the overall meta-analysis (Section 3.1.2), subgroups were inspected, this time comparing effect sizes to the mean weighted effect size for that particular subgroup. Only two samples were excluded, one from the Black subgroup (Weston et al., 2007) and one from the South America subgroup. Subgroup analyses were then repeated. Using random effects with separate estimates of τ^2, I found that ethnicity was not a statistically significant moderator, $Q_{\text{between}}(5) = 9.77$, $p = .082$, although the result is approaching significance. Utilising a pooled estimate of τ^2 produced a similar result, $Q_{\text{between}}(5) = 7.49$, $p = .187$.

Given these results, I carried out a meta-analysis for each subgroup in order to investigate ethnicity further, acknowledging that formal tests were only suggestive of a moderating effect but failed to reach statistical significance.

3.2.2. Separate meta-analyses for ethnicities

For each of the six broad categories of ethnicity/origin, I carried out a separate meta-analysis. The results are summarised in Table 1.

Table 1. A summary of the meta-analyses for the six categories.

Category	***N***	***k***	***Q***	**I^2 (%)**	***d***	**95% CI**	***p***
White	2849	9	23.42*	65.84	-0.07	[-0.24, 0.11]	.450
Pacific Islands	404	4	10.05*	70.15	0.05	[-0.31, 0.41]	.798
Black	1951	9	24.64*	67.53	0.06	[-0.14, 0.26]	.572
South America	696	25	42.17*	43.08	0.13	[-0.08, 0.35]	.222
Australian Aboriginals	238	3	2.63	23.90	0.13	[-0.16, 0.42]	.386
East Asia	487	8	6.38	0	0.26	[0.09, 0.43]	.002

N is the total sample size, k is the number of studies, or populations in this case, and *d* is the mean weighted effect size. *Q* and I^2 are measures related to the amount of heterogeneity in the group. * Significant at an alpha level of 0.05. Negative values of I^2 are set to zero (Higgins et al., 2003).

While the meta-analysis of all populations of skulls suggested that men have larger FWHR than women (although effect sizes less than 0.2 are considered small), the separate analyses for each ethnicity and geographic origin perhaps support a different interpretation. Only East Asian skulls show evidence of an effect (and even then, it is small), with no other categories suggesting sexual dimorphism. In fact, if these eight East Asian populations were excluded, the remaining populations as a whole no longer provide (statistically significant) evidence of a sex difference, $N = 7454$, $k = 79$, mean weighted $d = 0.07$, 95% CI [-0.01, 0.15], $p = .087$.

For completeness, the Black and South America subgroups were re-analysed after exclusion of the previously mentioned outliers (Section 3.2.1.). For the Black populations, the result remained qualitatively unchanged, $N = 1891$, $k = 8$, mean weighted $d = -0.02$, 95% CI [-0.19, 0.15], $p = .818$. For the South American populations, removal of the outlier produced an almost significant result, $N = 680$, $k = 24$, mean weighted $d = 0.18$, [-0.01, 0.37], $p = .067$.

Interestingly, many of the categories show moderate to large inconsistencies (I^2), perhaps suggesting the presence of further moderators or simply that the broad labels used here remain too inclusive and require additional subdivisions (see Section 2.3). In contrast, the East Asian studies showed no observed heterogeneity (and, as a result, provide values similar to a fixed effects model). This suggests that these particular samples are all measuring the same construct.

3.3. Reanalysis of Geniole et al. (2015)

In a previous meta-analysis, Geniole et al. (2015) found a small but statistically significant difference between the FWHR of men and women. However, the researchers included studies conducted on both faces and skulls, and did not discuss the possibility of differences between ethnicities. Previous evidence has shown that facial dimensions vary across ethnicities (Fang et al., 2011). I therefore reanalysed their data while taking into account the possibility of differences between faces and skulls, and the potential effect of ethnicity.

3.3.1. Skulls versus faces

The meta-analysis results for samples of faces, using the authors' unaltered data (Table S1 from Geniole et al., 2015) but *excluding* the eight samples which investigated skulls, found no (statistically significant) evidence of the presence of sex differences, $N = 4161$, $k = 24$, mean weighted $d = 0.12$, 95% CI [-0.01, 0.25], $p = .068$. Of course, although no longer significantly different from zero, this result remains similar to the original analysis with skull samples included, mean weighted $d = 0.11$, [0.03, 0.20]. Indeed, a subgroup analysis using random effects with separate estimates of τ^2 showed that source (skulls, faces) was not a statistically significant moderator, $Q_{\text{between}}(2) = 0.17$, $p = .920$. However, this analysis included samples of all ethnicities, which may be one reason why I found a large amount of within-group heterogeneity (I^2): skulls – 47%, faces – 75%.

3.3.2. Meta-analyses of White faces

For several of the studies of faces in the previous meta-analysis, the ethnicities of the participants were either unreported in the original papers or included a mixed sample. A meta-analysis of the studies with only White samples (the only category which included more than two studies) again found no evidence of the presence of sex differences in faces, N = 2037, k = 9, mean weighted d = -0.05, [-0.21, 0.11], p = .559. In this case, the result is very similar to that of White skulls (see Table 1).

3.4. Meta-analysis of Chinese faces

In skulls, only East Asians seemed to provide some evidence of larger FWHR in men (see Section 3.2.2). However, the majority of studies on FWHR in faces have focussed on White populations. Therefore, in order to investigate whether faces of this ethnicity demonstrate sex differences in FWHR, I carried out a meta-analysis that included two previous studies of Korean faces (Huh, 2013; Huh et al., 2014) and an additional sample of images that I had collected a few years ago. Front-facing photographs, with neutral expressions, were taken of 135 Chinese students (56 men; age range 19-44; age M = 23.15, SD = 3.30) at Bangor University, UK. In line with previous research, images were rotated so that both pupils were aligned to the same transverse plane, and then FWHR was calculated using the horizontal distance between the zygions, and the vertical distance between the highest point of the upper lip and the highest point of the eyelids (Kramer et al., 2012).

A summary of the three samples can be found in the Supplementary Materials. A meta-analysis of these face samples found no (statistically significant) evidence of the presence of sex differences, N = 331, k = 3, mean weighted d = 0.44, [-0.24, 1.12], p = .204. However, the point estimate (considered around a medium-sized effect) is noticeably higher than for the

White faces, and the confidence intervals include large effect sizes as well as zero. Of course, the inclusion of additional samples would provide a more precise estimate of the true effect.

4. Discussion

The current meta-analyses provide evidence that casts doubt on what seems to be the currently accepted story regarding FWHR dimorphism. Although an overall analysis of human skulls found a very small (though statistically significant) effect, where men showed larger FWHR than women, there is an argument to be made for considering ethnicities separately (Gill & Rhine, 2004; İşcan & Steyn, 1999; Ousley et al., 2009). Subgroup analyses were suggestive of a moderating effect of ethnicity (although these failed to reach statistical significance). However, after carrying out separate analyses for six ethnicities/geographical origins, I found that only East Asians demonstrated FWHR sexual dimorphism, and again, this effect was small, although notably larger than for the other groups. Given the between-population differences in skulls, it may be that some populations show sexual dimorphism for this ratio while others do not. Such a result goes against the idea that FWHR represents an evolved signal as a result of sex differences during development if we accept that differences are only (weakly) present in a limited number of populations.

The investigation of ethnicity presented here is necessarily limited by the availability of samples that can be reasonably pooled within subgroups. Given that most craniometric variation exists within local populations rather than between geographic regions (Relethford, 1994), it is important that further research investigates differences between specific populations (e.g., Han Chinese) where sufficient numbers of samples can be obtained. In this way, we should be better able to identify which populations demonstrate sexual dimorphism in FWHR and which do not.

Of course, it may be that skulls do not show sexual dimorphism with regard to FWHR, and instead, humans have evolved a signalling system based upon facial soft tissue deposits. Sex differences in soft tissue thickness may play a role in FWHR cues (Enlow, 1982; Lefevre et al., 2013), and previous evidence has established an association between FWHR and body mass index (Coetzee et al., 2010; Kramer et al., 2012). A recent meta-analysis found a "small but significant difference" in FWHR across 32 samples (Geniole et al., 2015, p. 14) where the majority of studies involved face measurements. Crucially, the researchers did not carry out a separate analysis for faces alone after excluding samples of skulls. (Also remember that their face and skull samples did not differentiate between ethnicities.) Here, I found that a replication of their analysis after excluding skull samples failed to identify a (statistically significant) difference between men and women. In addition, by controlling for ethnicity and considering only White populations (the ethnicity best represented), the effect decreased further to the point where there was no suggestion of a sex difference. However, mirroring the results with skulls, there may be some evidence suggesting an effect for East Asian faces, although more samples are needed before we can make any firm claims regarding the presence of FWHR dimorphism in specific ethnicities.

The first studies in this field provided strong evidence that men had larger FWHR than women for both skulls (d_{unb} = 0.84, 95% CI [0.32, 1.38]; Weston et al., 2007) and faces (d_{unb} = 0.50, [0.10, 0.96]; Carré & McCormick, 2008). In the intervening years, researchers have found mixed results, and I show here that there is no compelling evidence to support the initial hypothesis that FWHR is sexually dimorphic. Interestingly, the sample reported in Weston et al. was identified here as an outlying effect size in both the meta-analysis of all populations and in the Black subgroup analysis. While there is no general evidence of sexual dimorphism in the current work, there may be an exception for specific ethnicities or populations, although even in these cases, the effects remain relatively small. Statistically, it makes little sense to state with complete certainty that there is “no effect” (i.e., no difference between the FWHR of men and women), but I argue that consideration of the evidence presented here leads us to conclude that, at most, the effect is very small or absent.

How do we reconcile this conclusion with the growing evidence that FWHR is a reliable predictor of various behaviours (Geniole et al., 2015; Haselhuhn et al., 2015)? If we rule out the idea that FWHR cues are the result of sexual selection, through the exaggeration of a sexually dimorphic trait, then it is still possible that other mechanisms are responsible for the FWHR–behaviour association. However, the most likely contender was a testosterone-based mechanism but this has failed to find recent support in large samples (Bird et al., 2016; Hodges-Simeon et al., 2016).

Could we explain facial cues using a perception-based mechanism instead? There is very strong evidence that those with higher FWHR are perceived to be more aggressive and dominant (Geniole et al., 2015). Perhaps this is because relatively wider faces subtly resemble angry expressions, and people’s perceptions are the result of an overgeneralisation of their judgements of emotional expressions (Said et al., 2009). Interestingly, angry faces do not actually have higher FWHR than neutral expressions (Kramer, 2016; although Marsh et al., 2014, find the opposite result when FWHR is measured differently).

A second possibility has been couched in terms of “babyfaceness” – having a rounder face, and as a result, a higher FWHR. Previously, evidence had shown that boys who appeared more babyfaced displayed higher academic achievement if they were motivated to do so, but if they came from lower socioeconomic backgrounds, they showed more criminal behaviours (Zebrowitz et al., 1998a). In addition, early babyfaceness was associated with assertiveness and hostility later in life (Zebrowitz et al., 1998b). This result was explained as a self-defeating prophecy, whereby boys compensated for the warm or naïve stereotypes that people applied to them by manifesting personality traits that counteracted expectations.

However, there is now evidence that childhood babyfaceness and infant FWHR are both associated with infant temperament (Arcus & Kagan, 1995; Zebrowitz et al., 2015). These researchers also found a significant correlation between babyfaceness and adult FWHR. Taken together, the suggestion is that a bolder temperament from a larger FWHR in infancy extends through life, resulting in babyfaced adults (who have larger FWHR) demonstrating more assertive and aggressive behaviours. In support of this idea, longitudinal studies have shown that infant temperament predicts behaviour in adolescence and adulthood (e.g., Schwartz et al., 2012). While the mechanism linking infant temperament and facial appearance remains unknown, possible candidates include prenatal glucocorticoid or testosterone exposure (Arcus & Kagan, 1995).

In conclusion, I find a lack of evidence suggesting FWHR differences between men and women, both in skulls and in faces. Considered alongside recent evidence that FWHR does not appear to be associated with testosterone, researchers should now seek new mechanisms in order to explain the relationship between FWHR and behaviour.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Arcus, D., & Kagan, J. (1995). Temperament and craniofacial variation in the first two years. *Child Development, 66*, 1529-1540.

Bird, B. M., Cid Jofré, V. S., Geniole, S. N., Welker, K. M., Zilioli, S., Maestripieri, D., … Carré, J. M. (2016). *Evolution and Human Behavior*. Advance online publication.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305-327.

Burton, A. M., Bruce, V., & Dench, N. (1993). What's the difference between men and women? Evidence from facial measurement. *Perception, 22*, 153-176.

Calcagno, J. M. (1981). On the applicability of sexing human skeletal material by discriminant function analysis. *Journal of Human Evolution, 10*, 189-198.

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*, 2651-2656.

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

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics: Estimation, open science, and beyond.* New York: Routledge.

Darwin, C. R. (1872). *The expression of the emotions in man and animals*. London: John Murray.

de Villiers, H. (1968). *The skull of the South African Negro: A biometrical and morphological study*. Johannesburg: Witwatersrand University Press.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*, 203-235.

Enlow, D. H. (1982). *Handbook of facial growth*. Philadelphia: Saunders.

Fang, F., Clapham, P. J., & Chung, K. C. (2011). A systematic review of inter-ethnic variability in facial dimensions. *Plastic and Reconstructive Surgery, 127*, 874-881.

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.

Gill, G. W., & Rhine, S. (Eds.). (2004). *Skeletal attribution of race: Methods for forensic anthropology*. Albuquerque, NM: Maxwell Museum of Anthropology.

Gómez-Valdés, J., Hünemeier, T., Quinto-Sánchez, M., Paschetta, C., de Azevedo, S., González, M. F., … González-José, R. (2013). Lack of support for the association between facial shape and aggression: A reappraisal based on a worldwide population genetics perspective. *PLoS ONE, 8*(1), e52317.

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), e0122637.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*, 557-560.

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

Howells, W. W. (1973). *Cranial variation in man: A study by multivariate analysis of patterns of differences among recent human populations: Vol. 67. Papers of the Peabody Museum of Archaeology and Ethnology.* Cambridge, MA: Harvard University Press.

Howells, W. W. (1989). *Skull shapes and the map: Craniometric analyses in the dispersion of modern Homo: Vol. 79. Papers of the Peabody Museum of Archaeology and Ethnology.* Cambridge, MA: Harvard University Press.

Howells, W. W. (1995). *Who's who in skulls: Ethnic identification of crania from measurements: Vol. 82. Papers of the Peabody Museum of Archaeology and Ethnology.* Cambridge, MA: Harvard University Press.

Huh, H. (2013). Digit ratios, but not facial width-to-height ratios, are associated with the priority placed on attending to faces versus bodies. *Personality and Individual Differences, 54*, 133-136.

Huh, H., Yi, D., & Zhu, H. (2014). Facial width-to-height ratio and celebrity endorsements. *Personality and Individual Differences, 68*, 43-47.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage Publications.

İşcan, M. Y., & Steyn, M. (1999). Craniometric determination of population affinity in South Africans. *International Journal of Legal Medicine, 112*, 91-97.

Jantz, R. J., & Moore-Jansen, P. H. (2006). Database for forensic anthropology in the United States, 1962-1991: Inter-university Consortium for Political Social Research (ICPSR) [distributor]; 2006.

Jones, A. L., Kramer, R. S. S., & Ward, R. (2012). Signals of personality and health: The contributions of facial shape, skin texture, and viewing angle. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 1353-1361.

Kramer, R. S. S. (2016). Within-person variability in men's facial width-to-height ratio. *PeerJ, 4*, e1801.

Kramer, R. S. S., Jones, A. L., & Ward, R. (2012). A lack of sexual dimorphism in width-to-height ratio in White European faces using 2D photographs, 3D scans, and anthropometry. *PLoS ONE, 7*(8), e42705.

Kramer, R. S. S., & Ward, R. (2010). Internal facial features are signals of personality and health. *The Quarterly Journal of Experimental Psychology, 63*, 2273-2287.

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*, 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*, 273-279.

Lestrel, P. E. (1974). Some problems in the assessment of morphological size and shape differences. *Yearbook of Physical Anthropology, 18*, 140-162.

Marsh, A. A., Cardinale, E. M., Chentsova-Dutton, Y. E., Grossman, M. R., & Krumpos, K. A. (2014). Power play: Expressive mimicry of valid agonistic cues. *Social Psychological and Personality Science, 5*, 684-690.

Montepare, J. M., & Opeyo, A. (2002). The relative salience of physiognomic cues in differentiating faces: A methodological tool. *Journal of Nonverbal Behavior, 26*, 43-59.

Ousley, S., Jantz, R., & Freid, D. (2009). Understanding race and human variation: Why forensic anthropologists are good at identifying face. *American Journal of Physical Anthropology, 139*, 68-76.

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*, 169-173.

Relethford, J. H. (1994). Craniometric variation among modern human populations. *American Journal of Physical Anthropology, 95*, 53-62.

Relethford, J. H. (2002). Apportionment of global human genetic diversity based on craniometrics and skin color. *American Journal of Physical Anthropology, 118*, 393-398.

Rhodes, M. G. (2009). Age estimation of faces: A review. *Applied Cognitive Psychology, 23*, 1-12.

Rightmire, G. P. (1970). Bushman, Hottentot and South African Negro crania studied by distance and discrimination. *American Journal of Physical Anthropology, 33*, 169-195.

Rule, N. O., Ambady, N., & Hallett, K. C. (2009). Female sexual orientation is perceived accurately, rapidly, and automatically from the face and its features. *Journal of Experimental Social Psychology, 45*, 1245-1251.

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion, 9*, 260-264.

Schwartz, C. E., Kunwar, P. S., Greve, D. N., Kagan, J., Snidman, N. C., & Bloch, R. B. (2012). A phenotype of early infancy predicts reactivity of the amygdala in male adults. *Molecular Psychiatry, 17*, 1042-1050.

Scott, N. J., Kramer, R. S. S., Jones, A. L., & Ward, R. (2013). Facial cues to depressive symptoms and their associated personality attributions. *Psychiatry Research, 208*, 47-53.

Sell, A., Cosmides, L., Tooby, J., Sznycer, D., von Rueden, C., & Gurven, M. (2009). Human adaptations for the visual assessment of strength and fighting ability from the body and face. *Proceedings of the Royal Society B: Biological Sciences, 276*, 575-584.

Smedley, A., & Smedley, B. D. (2005). Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. *American Psychologist, 60*, 16-26.

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*, 551-556.

Weston, E. M., Friday, A. E., & Liò, P. (2007). Biometric evidence that sexual selection has shaped the hominin face. *PLoS ONE, 2*(8), e710.

Zebrowitz, L. A., Andreoletti, C., Collins, M. A., Lee, S. Y., & Blumenthal, J. (1998a). Bright, bad, babyfaced boys: Appearance stereotypes do not always yield self-fulfilling prophecy effects. *Journal of Personality and Social Psychology, 75*, 1300-1320.

Zebrowitz, L. A., Collins, M. A., & Dutta, R. (1998b). The relationship between appearance and personality across the life span. *Personality and Social Psychology Bulletin, 24*, 736-749.

Zebrowitz, L. A., Franklin, R. G., Jr., & Boshyan, J. (2015). Face shape and behavior: Implications of similarities in infants and adults. *Personality and Individual Differences, 86*, 312-317.

Figure Captions

Fig. 1. Craniofacial landmarks used to calculate FWHR.
The skull width (the distance between the left and right zygions) is divided by the upper facial height (the distance between the nasion and prosthion) to produce the FWHR. Figure adapted from Weston et al. (2007).

Supplementary Materials

InformationOnDatasets.xlsx
These spreadsheets provide information on the populations included in the meta-analyses.

S1 Fig. Effect sizes for the 87 populations included in the skull meta-analysis.
The mean weighted effect size is highlighted in grey on the left. The eight outlying effect sizes are labelled with red arrows.