**Challenges to both reliability and validity of masculinity-preference measures in menstrual-cycle-effects research.**

Michael B. Lewis

School of Psychology, Cardiff University, Park Place, Cardiff, CF10 3AT, UK.

Word Count: 8,225

Tables: 1

Figures: 6

Highlights

Typical measures of masculinity preference are confounded with accuracy and so have poor validity.

Typically, small numbers of preference decision leads to poor reliability in fertility preference studies.

No fertility shift in masculinity preference is found when validity and reliability of the measures are improved.

Face matching performance is better at times of higher fertility, which can be mistaken for a shift in preference.

Keywords: Menstrual cycle; Mate preferences; Masculinity; Racial features; 2afc preferences.

Abstract

Although it remains contentious, women's changeable attraction to masculine faces has been used to inform evolutionary ideas about human mating strategies. Typical experiments in this area use two-alternative-forced-choice (2afc) over a few pairs of similar images. The reliability of these measures is analysed suggesting that many studies have too few trials to be reliable. In the current experiment, fertility shifts in preferences for masculinised faces (and Africanised faces) were explored using both attractiveness ratings and a 2afc method over 80 pairs. The 2afc method showed a fertility shift in preferences whereas attractiveness ratings did not show a shift. Further, it was demonstrated how the size of the preferences shown in the 2afc tasks correlated with general face-matching performance. It is concluded that fertility is associated with improved face-processing accuracy and hence 2afc designs have poor validity as measures of masculinity preference. These issues of validity and reliability may have contributed to the contentious nature of fertility effects on preferences. Further, validity and reliability need to be considered in any study where a change in preference is identified using a comparative-preference task.

**Challenges to both reliability and validity of masculinity-preference measures in menstrual-cycle-effects research.**

One highly productive area of evolutionary psychology has been the attempt to understand whether, and how, women's preferences for men change as fertility changes over the menstrual cycle. The evolutionary premise of this research is that mate preferences that are increased at times of high fertility are going to be for features that indicate inheritable genetic quality that are desired for the offspring. Those preferences in a mate that are increased at times of low fertility are those features that are associated with the potential for long-term caregiving (Gildersleeve, Haselton & Fales, 2014a; Penton-Voak, Perrett, Castles, Kobayashi, Burt, Murray & Minamisawa, 1999). In support of this, it has been demonstrated that women's preference for a dominant body-odour is increased at times of higher fertility (Havlicek, Roberts & Flegr, 2005) and also the preference for a taller mate is stronger at times of higher fertility (Pawlowski & Jasienska, 2005). Increased preferences for more masculine voices (Puts, 2005; Feinberg, Jones, Law Smith, Moore, DeBruine, Cornwell, Hillier, Perrett, 2006) and masculine body shapes (Little, Jones & Burriss, 2007) have been linked to levels of increased fertility as has preference for facial symmetry (Little, Jones, Burt & Perrett, 2007; Oinonen & Mazmanian, 2007). Further, increased fertility has been shown to be related to increased preference for cuter baby faces (Lobmaier, Probst, Perrett & Heinrichs, 2015) and men with larger pupils (Caryl, Bean, Smallwood, Barron, Tully & Allerhand, 2009). One of the main topics in this field, and the one focused on here, has been the examination of whether the preference for masculine faces is increased during the fertile stage over the non-fertile stage of the menstrual cycle (e.g., Penton-Voak et al, 1999).

Research into menstrual-cyclic shifts in preferences has been dominated by the use of comparative measures of preference: decisions as to which face in a set is more attractive. Here, it is demonstrated that these comparative measures of preferences can be poor in both their validity and their reliability. An experiment is presented in which preferences for masculine faces are evaluated across the menstrual cycles using measures that improve both the reliability and validity of the level of masculinity preferences. While the research focuses on masculinity preference it also explores

racial preference, but its implications could apply to any preference task where the fertility cycle has been shown to produce a shift in preferences.

**Fertility based shifts in masculinity preferences**

Most of the research that investigates fertility-based shifts in preferences uses comparative measures. These methods involve making judgements of comparative attractiveness of items presented either simultaneously or consecutively while these items vary on only one dimension. The most common of these methods is the two-alternative-forced-choice (2afc) design whereby the participant chooses the preferred item from two simultaneously presented items. Alternative comparative measures have involved more than two alternatives in the forced choice or the use of a slider that gradually changes an image until the ideal face is selected. The use of a slider is still a comparative methodology because the participant is required to make a series of preference decisions over consecutively presented faces. By allowing the participants to compare very similar faces, these comparative measures mean that it is possible to obtain a measure of preference based just on the manipulated dimension using very small numbers of comparisons (typically 10 or fewer).

The first demonstration of the menstrual cycle affecting masculinity preference was a report in *Nature* showing that, at times of lower conception risk, the preference for feminised faces was greater than at times of higher conception risk (Penton-Voak et al, 1999). The participants' task was to select the most attractive from a set of five faces that varied only on their masculinity. While the fertility effect was present in Experiment 1 (more masculine faces were selected during times of higher fertility), it disappeared in Experiment 2 and was replaced by an interaction with length of relationship. A similar five-face-choice design was repeated with a larger sample by Penton-Voak & Perrett (2000). In this study, those in the low fertility part of their cycle selected each of the five faces with approximately equal frequency whereas those in the high fertility part of their cycle selected the slightly masculinised face more than any of the other faces therefore showing a stronger masculinity preference. In a much larger study, Jones, Little, Boothroyd, DeBruine, Feinberg, Law Smith, Cornwell, Moore & Perrett (2005a) used a 2afc task to assess preference between a masculinised or feminised version of faces. They found that women selected the masculinised face more often than the

feminised face when they were in a fertile phase (late-follicular phase) than in a non-fertile phase (mid-luteal phase). A similar 2afc procedure was used by Little and Jones (2012) and found (but only for short-term relationships) a stronger masculinity preference during the fertile phase (days 6-14) than the non-fertile phase (outside of these days). Studies using hormone measurements show that higher estradiol levels (found during the fertile phase) are associated with increased preference for masculine faces (Ditzen, Palm-Fischbacher, Gossweiler, Stucky & Ehlert, 2017) in 2afc tasks. A 2afc design using real faces presented in pairs with a high masculinity and low masculinity item have also shown a link between fertility and the strength of the masculinity preference (Little, Jones & Debruine, 2008). Stronger preferences for masculine faces at times of higher fertility have also been found in studies where participants used a slider to change the properties of a face (Johnston, Hagel, Franklin, Fink & Grammer, 2001; Vaughn, Bradley, Byrd-Craven & Kennison, 2010).

In spite of the many findings of the menstrual-cycle-related shifts in masculinity preferences, it remains controversial. For example, Harris (2010; 2013), in a large scale study found no link between menstrual cycle and masculinity preference. The task involved selecting the preferred face from a set of 5 faces of varying degrees of masculinity.

Two meta-analyses have been carried out in attempts to bring order to the data. Gildersleeve et al. (2014a) found a near-significant overall effect of cycle on masculinity preference and a significant preference when only short-term-relationship contexts were analysed. Wood, Kressel, Joshi & Louie (2014), however, found that fertile women did not show a stronger preference for masculine attributes than non-fertile women and there was no effect of length of relationship. The differences in these two meta-analyses were addressed by Harris, Pashler and Mickes (2014) who indicated that Gildersleeve et al. (2014a) had not adjusted for the potential researcher freedom that there is in selecting fertile versus non-fertile periods. This freedom allows for selection of target ranges that are most favourable for the results leading to an increase in observed effect sizes (which would be an example of *p*-hacking). Gildersleeve, Haselton & Fales (2014b) provide a robust defence of their meta-analysis suggesting that the *p*-hacking could not be the sole reason for the observed shifts in masculinity preferences. They claim that Wood et al. overlooked clear evidence for the shift in masculinity preferences. Regardless of whether *p*-hacking did occur in past studies or not, the best

recommendation for future studies would be pre-registration (see Wagenmakers, Wetzels, Borsboom, van der Maas & Kievit, 2012) with a clear definition of fertile and non-fertile periods recorded prior to the research being carried out – a strategy that was employed in the current research.

Further studies have investigated the fertility effect on masculinity preference since these meta-analysis studies. Little & Jones (2012) and Dixson, Blake, Denson, Gooda-Vossos, O'Dean, Sulikowski, Rantala & Brooks (2018) found significant fertility shifts for masculinity preferences; however, some have failed to find a fertility shift (e.g., Marcinkowsha, Galbarczyk & Jasiensak, 2018; Zeitsch, Lee, Sherlock & Jern, 2015, and Jones, Hahn, Fisher et al., 2018). Jones, Hahn & DeBruine (2018) provide a recent review of the topic and some of the methodological issues associated with finding a fertility shift for masculinity preferences, although they do not consider the two main methodological issues raised here: that of the validity and reliability of the masculinity preference tasks typically employed.

Some studies into the effect of fertility on masculinity preference have used tasks that are not comparative preferences, such as obtaining attractiveness ratings for faces varying on their masculinity. Peters, Simmons & Rhodes (2009) specifically explored the correlation between attractiveness ratings and masculinity across the menstrual cycle and found no relationship between fertility and the degree preference for more masculine faces. Further, although it was not the main focus of their studies, both Bressan & Stranieri (2008) and Rupp, Librach, Feipel, Ketterson, Sengelaub & Heiman (2009) found that the difference in attractiveness ratings for masculine and feminine faces was unrelated to menstrual cycle. Gildersleeve et al.'s (2014a) meta-analysis also cites unpublished data and conference data that used attractiveness ratings and failed to link masculinity preference to fertility.

The importance of the nature of the task used to assess masculinity preference can be demonstrated by re-examining the data from Gildersleeve et al.'s (2014a) meta-analysis. These data were reanalysed looking at the moderating effect of comparative versus ratings-based measures of attractiveness for those studies that looked at facial masculinity (all other moderating factors, such as length of relationship, were ignored). The overall effect size for just the comparative-measures experiments was significant, Hedges' $g = 0.21$, $p < .0001$, or Hedges' $g = 0.25$, $p < .001$ for just

alternative-choice designs, whereas the overall effect size for just the ratings-based experiments was not significant (and in the opposite direction), Hedges' $g$ = -0.15, $p$ = .089. The moderating effect of type-of-design (comparative versus ratings-based) was significant, $z$ = 3.39, $p$ < .001. This suggests that using comparative designs greatly increases the chances of finding a fertility effect on masculinity preference; however, this does not mean that comparative measures are the best way to assess masculinity preference as discussed below.

**The validity of comparative measures of preference.**

Here it is suggested that comparative measures of masculinity preference have poor validity and are potentially testing facial-processing ability rather than masculinity preference. The problem of validity arises because the measure is simply, in the case of a 2afc design, the proportion of times that a more masculine face is selected over a more feminine face. The typical assumption is that a change in masculinity preference would mean that it is more likely that the masculine face will be selected from the pair. However, the number of times the more masculine face is selected could be influenced by more mundane cognitive aspects such as levels of concentration on the task or general visual processing ability. If a person would normally select a more masculine face with a probability of more than .5 then any deficit in the ability to process faces or disengagement with the task would lead to a regression to the mean and a reduction in the number of masculine faces selected. This would be interpreted as a reduction in masculinity preference whereas it is a reduction in face processing ability that has nothing to do with the underlying preference for masculine faces.

There are good reasons to think that the menstrual cycle could have a direct effect on facial-processing ability. In general, visual sensitivity is higher at times of ovulation compared to the premenstrual phase (Friedman & Maeres, 1978; Parlee, 1983). More specifically, brain imagining supports the link between female hormones and differences in face processing because activation of the right fusiform face area has been found to be greater in the follicular phase than during menses (Marečková, Perrin, Khan et al. 2014). Also, directly administered progesterone has been found to reduce face-recognition accuracy (van Wingen, van Broekhoven, Verkes, Petersson, Bäckström, Buitelaar & Fernández, 2007).

The validity of non-comparative measures of preference, for example where faces of varying masculinity are rated for on their attractiveness (e.g., Peters et al., 2009), can also be considered. In these studies it is the strength of the correlation between attractiveness ratings and masculinity that is used to infer the level of preference for masculine faces. If there is a shift in face-processing ability then this would introduce extra noise into the measurements and hence reduce the correlation. This can explain why Debruine (2012) observed a correlation between masculinity preference based on ratings and masculinity preferences based on a 2afc task. While this is still a threat to the validity of the measure of preference, this threat is smaller than for comparison-based measures and can be reduced further by using larger sets faces. Probably the most valid measure would be to obtain attractiveness-ratings-by-masculinity curves and look at the properties of those curves such as maximums. This would be time consuming, but it is a measure that has been used effectively to assess individual differences in preferences (Holzleitner & Perrett 2017) but not yet for menstrual-cycle effects. The solution to the validity problem employed here was to assess masculinity preferences using both ratings and a 2afc task but then partial out any variation in the abilities to distinguish between very similar faces.

This issue of validity of preference measures was at the centre of the study by Lewis (2017) in which comparative attractiveness preferences were compared to rating-based assessments of attractiveness when assessing the role fertility has on the preference for symmetrical faces. Lewis (2017) showed that there was a larger preference for symmetrical faces during fertile periods when a 2afc task was employed; however, when attractiveness ratings were assessed, there was no increased preference for symmetrical faces at times of increased fertility. The largest correlation with fertility was on a task where participants had to simply match faces and not make any judgements of their attractiveness. Increased face-matching ability correlated with the size of the symmetry preference in the 2afc task. The conclusion was that the fertility-based change to symmetry preference observed in the 2afc task in that experiment, and possibly previous experiments, were a result of changes in general face-processing ability rather than a change in preferences for symmetrical or asymmetrical faces. In the current experiment, a similar validity evaluation is made for 2afc measurements of masculinity preferences.

A fertility-related shift in face processing ability cannot explain all the fertility-related face preference shifts observed in the literature. For example, Penton-Voak et al. (1999) and Jones et al. (2005a) both showed that an overall preference for feminised faces was reduced at times of higher fertility showing. These studies are discussed further with regards to their reliability below, and also discussed in the context of the current experiment in the discussion.

**The reliability of comparative measures of preference.**

One methodological issue that may explain why some studies find a masculinity shift and others do not may be the small number of items typically used in this type of research: Penton-Voak et al. (1999) used five trials; Jones et al. (2005a) used three trials, and both Little & Jones (2012) and Jones and colleagues (2018) used ten trials in each condition. There are two potential problems of using such small numbers of trials. First, the results could be peculiar to the images selected and so there is a potential lack of generalizability. This would be similar to the language-as-a-fixed-effect fallacy (Clark, 1973). Second, a more troublesome problem with small numbers of trials is the lack of reliability of the data obtained as explained below. This lack of reliability undermines the power of those experiments.

Data modelling can be used to show how small numbers of binary decisions, as in 2afc designs, yield unreliable estimates of the underlying performance. Let us assume that actual masculinity preference levels are somewhere between .60 and .70 (as in the proportion of times that someone would prefer a masculine over a feminine face and so .50 means there is no masculinity preference). That is, each participant looks at a pair of faces and has a particular degree of preference for masculinity and that produces a probability of picking the more masculine face as being more attractive. We can use data modelling of a binomial distribution to investigate the expected correlation between actual preference and observed preference. One million data points were generated according to the distribution $B(10, p)$ where $p$ was allowed to vary uniformly between .60 and .70. In this simulation, $p$ represents the participant's actual masculinity preference and the random number generated from the distribution $B(10, p)$ represents their performance on a 10 item 2afc experiment. Over the one million data points, the correlation of each $p$ and the value $B(10, p)$ was $r = .19$:

suggesting poor reliability. The situation can also be modelled in which each participant is tested on 80 items in a 2afc experiment. In this case the values of $p$ are correlated against the randomly generated values using the distribution $B(80, p)$. In this case the correlation is $r = .48$, which is a much improved level of reliability. This clearly demonstrates the importance of number of items on reliability of a measure.

Changes in the reliability of the measure of masculinity preference will have consequences for the predicted sample size required to detect changes. Jones and colleagues (2018) suggest that very large numbers of participants are required to determine whether there are changes in masculinity preference. They employed 598 participants who assessed preferences using 2afc designs over sets of just 10 items and found no compelling evidence of a change in preference. As shown above, using just 10 items provides poor reliability and increasing the number of items to 80 items will increase the reliability of the measure of masculinity preference.

The exact impact that the number of items has on experimental design can be calculated by simulating the existing results. Assuming an effect size of Hedges' $g = 0.25$ (taken from the meta-analysis using only alternative-choice studies), a power analysis (alpha = .05, beta = .8) suggests 506 participants are required to investigate a fertility shift on masculinity preferences. If this effect size is based on a standard 10 item 2afc design (this is a little generous as the mean number of trials in forced-choice designs in the meta-analysis was 4.76) and we anchor fertile masculinity preference choices at $p_f = .70$ then using the fact that the data follow a binomial distribution we would expect non-fertile masculinity preference to be at $p_{nf} = .66$. That is the value $p_{nf}$ would have to take for the difference between $B(10, p_{nf})$ and $B(10, p_f)$ to have an effect size of 0.25. Therefore, we would hypothesise a difference of .04 in the masculinity preference between the fertile and non-fertile participants when expressed as a probability of selecting the more masculine face in the pair. Given this expected difference in the level of preference, we can determine the effect size if a 2afc design were used with 80 items instead of just 10 items: that is the effect size if we were comparing $B(80, p_{nf})$ and $B(80, p_f)$ with the same hypothesised preferences. The effect size in this case would be Hedges' $g = 0.71$ and a power analysis suggests that just 66 participants would be required for the experiment. This demonstrates how an improvement in the reliability of the measure can have a practical impact

on the experimental design. Putting it more simply, if you are looking for a difference of about .04 then a measure with a resolution of .10 (10 items 2afc design) is not going to be nearly as good as a measure with a resolution of .0125 (80 items 2afc design). The current experiment used 80 items in its 2afc designs suggesting 66 participants would be required; however, the actual sample size was determined using a Bayesian stopping rule (Rouder, 2014).

**Fertility and racial preferences in faces**

The main focus of the current research was to explain the apparent relationship between fluctuating fertility and preference for more masculine faces. However, if changing facial processing ability can explain the previous finding then preferences for any facial properties should show changes across the menstrual cycle. To investigate this, preference for racial characteristics were also assessed alongside masculinity in the current experiment.

Fertility effects on racial preferences have been investigated previously, although, with mixed results. Frost (1994) showed that the preference for darker faces increased during the fertile phase relative to the non-fertile phase. Further, Izbicki & Johnson (2010) found that White participants showed an increased preference for Black male faces in their fertile phase over their non-fertile phase. Contrary to these findings, Navarrete, Fessler, Fleischman & Geyer (2009) found a preference for White faces was even stronger during periods of higher fertility. In the current experiment, preferences for faces along a black-white continuum were investigated using both comparative and ratings methods.

**Experiment**

The current study investigated whether visual sensitivity changes associated with the menstrual cycle (e.g., Parlee, 1983) are driving the fertility-related changes in preference for masculinity (Little & Jones, 2012) and racial features (Frost, 1994). To do this, participants were tested in a cross-sectional manner on three different tasks within the same session (a fourth task provided a manipulation check for the stimuli). These tasks assessed facial preference using simple ratings and facial preferences using a 2afc design, and there was also a test of facial-processing

sensitivity. The preferences that were assessed were for both masculinity and for changes in the racial appearance of the faces. The study explored how performance on each of these tasks changed with changes in fertility but also, and more importantly, the correlations between these tasks were investigated to identify whether changes in performance of the 2afc tasks were related to changes in general face processing. To avoid the allure of *p*-hacking, the design was preregistered on the Open Science Framework https://osf.io/cnjkb/. The selection of test periods (nominally called fertile and non-fertile) was selected based on previous studies. As fertility is only loosely associated with day of menstrual cycle (Gangestand, Haselton, Welling, Gildersleeve, Pillsworth, Burriss, Larson & Puts, 2016), it is more likely that the any observed differences are a feature of hormone variations rather than fertility itself. However, consistent with the previous literature, we explore fertility effects rather than menstrual cycle effects, although the latter might be a better name for any findings. All stimuli, control files and datasets of available at https://osf.io/s59zf/.

**Method**

**Participants**

Participants were 93 female undergraduates. The aim was to recruit only participants that were not using oral contraceptives and had regular menstrual cycles by directing ineligible participants to a different study; however, due to an error in the other experiment, the full dataset does include some of these participants although their data do not form part of the current analysis. The number of participants was determined using a Bayesian stopping rule (Rouder, 2014, see below for the criteria). Participants' consent was obtained according to the Declaration of Helsinki and the protocol was approved by the local research ethics committee.

**Stimuli**

A set of 80 base faces were generated using the *Facegen Modeller 3.5* software package. This generates rendered 3D facial images based on the parameters collected from a large corpus of faces. The base faces were constructed such that they had a fixed set of demographic parameters but facial features were allowed to vary around these parameters in a natural manner. The fixed parameters were for: age, set at approximate aged 28 years; distinctiveness, set at the 'Typical' level; asymmetry, set at the 'Typical' level of asymmetry; and gender. The gender scale has 81 values ranging from 'very

female' to 'very male' and the base face took the 57th value on this scale. The last fixed parameters related to the 'race' of the faces and the faces were generated based on standard European features (this corresponded to points 31 on a 41 point scale for the racial scales with the upper end points being European). The angle, lighting conditions and the hairstyle were standard for all the faces.

From each of these base faces, five further faces were generated. First, the masculinity of the face was increased by 10 points on the scale to create a masculinised face. Second, the masculinity of the face was decreased by 10 points on the scale to create a feminised face. Third, the racial features of the face were increased beyond typical European features on the European-to-African scale by 7 points to create a hyper-European face. Fourth, the racial features on the European-to-African scale were moved towards the African features by 7 points to create an Africanised face. Lastly, the facial features were 50% morphed with a randomly created face to create a similar-looking yet different face referred to as the morphed face. See Figure 1 for a set of example images.

**Procedure**

Participants completed a short questionnaire concerning their contraception use, and the timing of their last period. Following this, participants then carried out a series of computer-based tasks using the faces described above.

***Task 1 – attractiveness ratings*.**

A series of 320 faces were presented individually to participants in a random order. These faces were the masculinised, feminised, hyper-European and Africanised faces for each of the 80 base faces. For each face, the participant indicated how attractive they thought the face looked. The scale went from 1 'least attractive' to 9 'most attractive'. Participants could take as long as they liked to make a response and the next face appeared after they made a response.

***Task 2 – 2afc attractiveness preference.***

In this task, 160 pairs of faces were presented in a random order. These pairs were either the masculinised and feminised faces or the hyper-European and Africanised faces versions of the same base face. For each base face, both pairs were used once (position of the two faces were counterbalanced across items). The participants selected which face in the pair that they thought was more attractive before the next pair was presented.

***Task 3 – face discrimination.***

The pairs of faces used in Task 2 were added to by pairs of faces that were the base face and the morphed face making 240 pairs. In this task, one member of the pairs of face was presented as a target face for 700ms followed by a fixation cross for 300ms. Immediately after, the target face and its other pair were presented at the same time. The participants' task was to identify whether the target face was on the left or right. Once a selection was made, a new target face was shown.

***Task 4 – masculinity rating.***

The series of 320 faces used in Task 1 were presented individually to participants in a random order. For each face, the participant indicated how masculine they thought the face looked. The scale went from 1 'least masculine' to 9 'most masculine'. Participants could take as long as they liked to make a response and the next face appeared after they made a response.

**Design**

The current study was designed to explore differences in face processing and face preferences for women in their fertile period compared with their non-fertile period. In line with previous research that has shown these correlations, the fertile period was defined as days 6 to 14 inclusive with non-fertile days being in the range day 19 onwards. A count forward method was used to assess the day of cycle. Participants were not included in the analysis if: they were using an oral contraceptive; were pregnant or had typical periods over 31 days or under 26 days.

A range of preference and performance measures were collected from the 4 tasks for each participant. The first four of these measures assessed preferences. From Task 1, the relative preference for masculinised faces over feminised faces was obtained. For each participant, this was the difference in mean ratings of the two sets of faces divided by the standard error of that participant's attractiveness ratings. Also from Task 1, the relative preference for Africanised faces over hyper-European faces was obtained in a similar way. Preferences were also generated from Task 2. The masculinised over feminised face preference was indicated by the number of times the masculinised face was selected as more attractive in the pair. The maximum was therefore 80 and a score of 40 shows no masculinised face preference. A similar 2afc-Africanised face preference was obtained from

the number of times, out of 80, that the Africanised face was selected as more attractive than the hyper-European face.

From Task 3, a measure of face-matching accuracy was obtained, which was the number of times the participant correctly identified the target face in the group of two. The measure was recorded as a score out of a maximum of 240 trials with 120 representing chance performance. This measure could be further broken down into three measures of performance distinguishing faces based on masculinity, racial features or random features; however, as described below, this level of fine grain analysis did not reveal any additional information and so the global measure in reported here only.

The purpose of Task 4 was a manipulation check. This was designed to ensure that the faces did differ in their perceived masculinity as would be suggested by the software that was used to create the stimuli.

Rather than a power analysis, a Bayesian stopping rule was employed to determine the number of participants. Participants were recruited, in groups of between 3 and 10, for the experiment until a set of criteria were met based on the Bayesian analysis of the data. The criteria were the fertile and non-fertile groups showed Bayes factors ($BF_{10}$) of either over 3.0 or under 0.333 for differences in the face-matching task and at least two of the attractiveness preference measures.

**Results**

Of the 93 participants tested, 9 were discounted because they either had irregular menstrual cycles or were using oral contraceptives. Twenty six were on days 6-14 (labelled here the fertile phase) whereas 36 were on days 19 plus (labelled here the non-fertile phase). The remaining participants fell outside of these two ranges. The following analyses were conducted on the 62 participants in the fertile and non-fertile phases only; however, the entire dataset is available as an open access resource.

**Task 1, Attractiveness ratings**

The average rated attractiveness for masculinised faces (mean = 3.47, standard error = 0.15) was higher than for feminised faces (mean = 3.11, standard error = 0.15) and the average rated attractiveness for Africanised faces (mean = 3.12, standard error = 0.16) was higher than for hyper-

European faces (mean = 2.64, standard error = 0.14). For each participant, standardised preference scores were calculated: For masculine preference, this was the difference between average attractiveness rating for masculinised faces and the average attractiveness rating for feminised faces divided by the standard error of the attractiveness ratings. For Africanised preference, this was the difference between average attractiveness ratings for Africanised faces and the average attractiveness ratings for hyper-European faces divided by the standard error of the attractiveness ratings. Figure 2 shows how these preferences differed between the two fertility groups. There was a slightly stronger masculinity preference in the non-fertile phase than the fertile phase although this was not significant and Bayesian analysis provides moderate evidence for the null hypothesis, $t(60) = 0.56$, $p = .57$, $BF_{10} = 0.30$. Further, there was a stronger Africanised preference in the fertile phase than the non-fertile phase but again this was not significant although the Bayesian analysis was inconclusive, $t(60) = 1.75$, $p = .085$, $BF_{10} = 0.93$. In this task, changes over the menstrual cycle are not linked to changes in preference for masculinised faces or Africanised faces.

**Task 2, 2afc attractiveness preference**

The average number of times that participants selected a masculinised face, over a feminised face, in a 2afc decision was 59.1 (standard error = 1.35) out of a possible 80 pairs. The average number for Africanised preferences over hyper-European faces was 58.1 (standard error = 1.76) out of 80 pairs. Figure 3 shows how these preferences differed between the two fertility groups. There was a stronger masculinity preference in the fertile phase than the non-fertile phase and this difference was significant and the Bayesian analysis provides moderate evidence for the hypothesis, $t(60) = 2.55$, $p = .014$, $BF_{10} = 3.71$. Further, there was a stronger Africanised preference in the fertile phase than the non-fertile phase and this was significant and the Bayesian analysis provides moderate support for the hypothesis, $t(60) = 2.47$, $p = .016$, $BF_{10} = 3.21$. In this task, changes over the menstrual cycle are linked to changes in preference for masculinised faces and Africanised faces.

**Task 3, Face-matching task**

The participants selected the correct target face in the subsequent pair, on average, 214.9 times (standard error = 1.50) out of a possible 240. Figure 4 shows how performance differed between the two menstrual cycle groups. Performance was better in the fertile phase than in the non-fertile

phase and this difference was significant and the Bayesian analysis provides moderate evidence for the hypothesis, $t(60) = 2.86$, $p = .006$, $BF_{10} = 7.24$.[1] Two participants in the non-fertile group performed particularly badly scoring just 172 and 180 out of 240. The overall results suggest that the visual discrimination or memory for faces is better in the fertile part of the cycle than in the non-fertile part of the cycle.

**Task 4, Masculinity ratings**

The average masculinity ratings for the four types of faces are shown in Table 1. The masculinity ratings for the masculinised faces were significantly higher than for the feminised faces, $t(61) = 7.44$, $p < .001$, which provides a manipulation check for the properties of the images. The masculinity ratings for the Africanised faces were also significantly higher than for the hyper-European faces, $t(61) = 7.34$, $p < .001$. This means that the racial change was confounded with a masculinity change.

**Inter-task correlations**

The correlations between the different tasks provide a greater insight into the mechanisms underlying the links between fertility, performance and preferences. The simple correlations are presented in Figure 5. Of the four measures of preference (two for masculinity and two for racial features), the only significant correlations where between the 2afc preference for Africanised faces and both rating-based preference for Africanised faces and 2afc preference for Masculinised faces. More generally, it can be observed that the correlation between the sizes of preferences shown in the 2afc tasks is larger than the correlation between the sizes of preferences shown in the rating-based tasks, $z = 3.72$, $p < .001$ (using *cocor* analysis for non-overlapping dependent correlation, Diedenhofen & Musch, 2015). The 2afc tasks, therefore, shows more within-participant consistency than the preferences based on ratings.

In order to address the questions of the relationship between menstrual cycle and preferences, two path analyses were conducted to compare the direct effect of fertility on 2afc preferences

[1] A mixed-ANOVA was also conducted with the added factor of type of comparison (masculinity change, racial change or morph change). While performance on the morph change was significantly worse than the other two changes, $F(2,120) = 57.5$, $p < .001$, (indicating that this change was a smaller visual change) there was no interaction between the type of comparison and menstrual cycle groups, $F(2,120) = 0.935$, $p = .395$.

compared with effects that can be seen as being moderated by general face processing accuracy. Figure 6 shows these two path analyses. For the 2afc preference for masculinised faces, the inter-participant variability is mostly predicted by the variability in the ability to match faces. The correlation between fertility and 2afc preference is largely mediated by the face matching ability. The pattern is similar for the 2afc preference for Africanised faces, with the fertility effect being moderated by the ability to match faces; however, in this case there is a link between rating preferences and 2afc preferences that is independent of both fertility and accuracy in the face matching task.

**Discussion**

This study assessed menstrual-cycle-based shifts in preferences for masculinity and for racial features using both a comparative task and a ratings-based task. These two tasks reveal different shifts, but these differences can be reconciled by looking at the performance in the face-matching task as described below.

The results from Task 2 give insights into 2afc preferences and how these change across the menstrual cycle: the typical method employed in previous research used to understand the shift masculinity preference. The results show that there is an increased 2afc preference for masculine faces during fertile periods, which is consistent with some previous research (Little & Jones, 2012) but this is a finding that has been difficult to find in recent studies (see Jones et al. 2018). One potential reason why the menstrual cycle effect was found in this study and not in some of the more recent studies could be the improved reliability of dependent variable. As described above, increasing the number of comparisons to 80 provides a more reliable measure of an individual's actual performance on the task than just using 10 comparisons.

Task 2 also found that there was an increased 2afc preference for Africanised faces during fertile periods. This finding is consistent with the findings of Frost (1994) but contrary to Navarrete et al (2009). If this were the only element of the investigation (which is typically the case in this kind of research) then the conclusions would be that women change their preferences as an attempt to maximize the positive inheritable traits from a mate during times of higher fertility: These positive inheritable traits being indicated by a more masculine face and also a more African-looking face.

Exactly why looking more African is a positive inheritable trait would have to be explained with some new and potentially provocative ideas concerning evolutionary psychology. However, this was not the only part of the research and further evaluation provides a different interpretation of the current findings and possibly of previous findings.

Task 1 looked at the attractiveness ratings for the masculinised, feminised, Africanised and hyper-European faces. Masculinised faces received higher average attractiveness than feminised faces and Africanised faces received higher average attractiveness than hyper-European faces, which was consistent with the racial differences observed by Lewis (2011). From the data, it was possible to obtain standardised scores for individual preferences for masculinised faces and individual preferences for Africanised faces for each participant. The size of the masculinised face preferences did not correlate with fertility, a finding that is inconsistent with the results of Task 2 but it is consistent with the findings of Peters et al. (2009) who used similar comparisons of attractiveness ratings. Similarly, the size of the Africanised face preference did not correlate with fertility which is again inconsistent with the results of Task 2. Taking Tasks 1 and 2 together, it appears that whether fertility is correlated with preferences is dependent on the nature of the task.

The purpose of Task 3 was to evaluate the general face-processing abilities of the participants. Lewis (2017) had shown that performance on this kind of face-matching task was associated with phase of menstrual cycles. This association was confirmed here with the finding that the fertile group had better performance on this task than the non-fertile group. This finding is consistent with the idea that changes over the menstrual cycle are associated with changes in visual sensitivity (e.g., Parlee, 1985; Friedman & Maeres, 1978) and these low-level effects could potentially affect higher-level visual decisions such as relative attractiveness. Given that some participants in the non-fertile group did particularly poorly on Task 3 it is possible that these participants were not paying as much attention during the experiment as other participants. Obviously, in this kind of task it is difficult to distinguish between differences in ability and differences in motivation.

How these changes in general face-processing ability could explain the different pattern of preferences found in Tasks 1 and 2 was explored using correlational analysis. The accuracy in Task 3 had low levels of correlations with the ratings-based preferences from Task 1 suggesting that these

preferences are fairly independent of the general face-processing abilities. The 2afc preferences obtained in Task 2 did, however, correlate with the face-matching accuracy in Task 3 such that those who showed greater masculinity preferences also showed greater accuracy in the face-matching task. This pattern of results supports the idea that fertility affects some low-level visual processing system and higher-level decisions that are underpinned by this system are similarly affected by fertility, or the non-fertile participants were less motivated to perform well during the experiment. Distinguishing the difference between two faces is a similar task regardless of whether the question being asked is "which face do you prefer?" as in Task 2 or "which face have you just seen?" as in Task 3. If one is better at or more motivated to seeing the difference between two faces then this would show as better performance in both Tasks 2 and 3. In this interpretation, it is clear why the path analysis shows that there is little direct effect of fertility on 2afc preferences once any mediating effect of face-processing ability is considered: the main direct effect that fertility has is on the face-matching task.

**Assessing the use of 2afc tasks more widely**

A 2afc preference task cannot disentangle differences in preferences and differences in the ability to distinguish between two items. In the current study, the apparent effects of menstrual cycle on masculinity preference are possibly a direct and artefactual consequence of using a 2afc to assess preference. It is possible that the same is true of other studies that have employed 2afc to assess variations in preferences for masculinity (e.g., Little & Jones, 2012). The finding that women prefer more African-like faces during times of higher fertility could have led to interesting speculation as to the evolutionary reasons behind this. Given that this association between preference for Africanised faces and fertility only appears in the 2afc task, the finding can be dismissed as an artefact. Indeed, the current interpretation can explain the contrasting results of Navarrete et al. (2009) and the current fertility-shift in the 2afc racial-features preferences. Navarrete et al. found that a preference for White faces was enhanced during times of higher fertility. Here, it was shown that a preference for Black faces was enhanced during times of higher fertility. The explanation that can resolve this pattern of results is that any bias is reduced when participants are in their non-fertile phase.

The current research shows that the use of 2afc designs can lead to seductive but potentially misleading results in the analysis of preferences for masculinity and racial features. Lewis (2017) demonstrated similar issues with the variations observed for preferences for symmetry in faces. In fact, any assessment of preference that is based on a 2afc decision could be prone to this kind of difficulty of interpretation. This analysis may explain why fertility has been shown to change masculinity preferences in body shapes when using comparative tasks (Little, Jones & Burriss, 2007) but not when using rating tasks (Jünger, Kordsmeyer, Gerlach, & Penke, 2018). Also, this analysis can explain why factors that are unrelated to reproduction appear to be enhanced at periods of fertility.

For example, Lobmaier et al (2015) showed that women's preference for cuter babies was stronger when they were in their fertile phase than in their non-fertile stage. The assessment for preference for cuteness involved a comparative 2afc design. Overall, cuter babies were selected more often but this preference was stronger during the ovulation phase than during the luteal phase (and in fact this difference was strongest when distinguishing between the two faces was hardest). Based on the research here, this effect can be re-interpreted as those women in the ovulation phase have better general face-processing skills and so were better able to distinguish between the two highly similar faces than those women in the luteal phase. Using this interpretation, the finding does not say anything about *kindchenschema* and care-giving behaviour but rather just about how low-level visual processing is improved at times of higher fertility. This explanation is also consistent with the fact that a similar fertility effect was not found when simple ratings of cuteness (i.e., a ratings-based measure of preference) were used in an earlier study (Sprengelmeyer et al. 2013). However, Lobmaier, Sprengelmeyer, Wiffen & Perrett (2010) showed, using the same stimuli, that women showed greater accuracy in a cuteness decision, but men showed greater accuracy in an age decision. This finding could be explained by the current model if, on average, women engaged more with the task when asked about cuteness but men engaged more with the task when asked about age – one could posit evolutionary theories as to why cuteness is more relevant for women but age is more relevant for men but such speculation is not required to account for the pattern observed.

One last example concerns the preference for pupil size. In Caryl et al's (2009) study, the size of the 2afc preference for larger pupils was greater for women in the fertile phase than in the non-

fertile phase. This result is particularly interesting as there is no additional inheritable advantage from mating with someone with larger pupils. Pupil sizes are constantly changing and they are a better indicator of the current illumination than of the genetic quality of the person. The current research suggests that Caryl et al.'s findings reflect facial-processing ability rather than mating strategy. The observed increase in preference for larger pupils seen at times of higher fertility could simply be a result of more accurately determining that the pupil sizes differed between the pairs of images being presented in the 2afc procedure.

These examples illustrate that the potentially invalid interpretation of 2afc preferences is not restricted to just masculinity preferences. The increased performance in distinguishing stimuli could be driving the observed menstrual shifts in preference in a range of tasks. Hence, careful consideration is required to ensure that any observed shifts in preference, be they for odor, size, gait, cuteness, pupil size, race, BMI or anything else, are not just a consequence of a change in low-level visual processing.

Not all of the observed fertility effects on preferences can be discounted as being due to better facial processing during fertile periods. In some studies, a preference for feminine faces becomes weaker as fertility increases, which goes opposite to the explanation in terms of better facial processing during times of higher fertility. However, these studies are characterised by particularly low numbers of trials (e.g., Jones et al., 2005a, study 2: three trials, Penton-Voak et al. 1999: five trials) and so the reliability is low in these studies. Other studies show that the strength of other preferences can go up during times of lower fertility. For example, DeBruine, Jones and Perrett (2005) found a stronger preference for self-resemblances during non-fertile days than fertile days using a comparative-preference task and Jones and colleagues (2005b) found that there was a stronger preference for apparent health during times of lower fertility. These cases (that show reduced preference at times of higher fertility) go against the current explanation, but Peter Hancock (personal communication) suggested that they could be a result of the artificial nature of the stimuli used. For example, masculinising a male face or making the face look more healthy may introduce artefacts in the images that are unattractive under the greater inspection that is associated with fertility – particularly when two almost identical faces are presented together. It could be the weaker detection

of these artefacts that leads to the greater preferences for feminine faces and healthy faces seen by non-fertile participants in Jones et al. (2005a).

The reason for the interest in the associations between fertility and any kind of preference is that it potentially offers an insight into our evolutionary psychology. Preferences that people have when they are fertile are more likely to get passed down to future generations. The fact that many of these preferences might just be a reflection of improved face processing ability does not mean the finding has no consequences for evolutionary psychology. It appears that humans are designed such that at times of heightened fertility, women are better able to process faces and read social signals. From an evolutionary point of view, this heightened skill in face processing would be generally useful in that it would be easier to identify the correct mate if the lighting conditions were poor. What constitutes the correct mate, be him more masculine, more symmetrical or anything else, might be something that does not necessarily vary over the menstrual cycle.

**Conclusion**

The current study demonstrated that there are menstrual cycle changes in the way faces are processed. This is not too surprising given the many demonstrations of menstrual cycle correlations with visual processing (e.g., Parlee, 1985). The current research explored whether there was a link between fertility and preference for more masculine features and racial features. Previous research, using comparative-preference designs, had shown that women in their fertile stage show a stronger preference for masculine faces than in their non-fertile stage (e.g., Little and Jones, 2012). Other research which compared attractiveness ratings did not show this shift in masculinity preference associated with fertility (Peters, et al. 2009). In the research reported here it was both found that there was a fertility shift in masculinity preference and also preference for African-looking faces in a 2afc design but also there was no fertility shift based on attractiveness ratings. The performance in the 2afc designs was found to be correlated and predicted by performance on a face-matching task. It was concluded, therefore, that the shifts in preferences that were observed here, and elsewhere, in a comparative-preference designs are possibly based on changes in the face-processing ability of participants across the menstrual cycle rather than a real change in mate preference. Ultimately, this research shows that 2afc tasks can be misleading when attempting to assess preferences.

**Acknowledgements**

Thanks to Peter Hancock and two anonymous reviewers for their ideas and comments on earlier drafts.

**Data Availability**

Summary data used for the analysed are available as supplementary material. The entire raw dataset is available on the Open Science Framework https://osf.io/s59zf/.

**References**

Bressan, P., & Stranieri, D. (2008). The best men are (not always) already taken: Female preference for single versus attached males depends on conception risk. *Psychological Science*, 19(2), 145-151.

Caryl, P. G., Bean, J. E., Smallwood, E. B., Barron, J. C., Tully, L., & Allerhand, M. (2009). Women's preference for male pupil-size: Effects of conception risk, sociosexuality and relationship status. *Personality and Individual Differences*, *46*(4), 503-508.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4), 335-359.

DeBruine, L. M. (2012). Evidence versus speculation on the validity of methods for measuring masculinity preferences: comment on Scott et al. *Behavioral Ecology*, 24(3), 591-593.

Ditzen, B., Palm-Fischbacher, S., Gossweiler, L., Stucky, L., & Ehlert, U. (2017). Effects of stress on women's preference for male facial masculinity and their endocrine correlates. *Psychoneuroendocrinology*, *82*, 67-74.

Dixson, B. J., Blake, K. R., Denson, T. F., Gooda-Vossos, A., O'Dean, S. M., Sulikowski, D., ... & Brooks, R. C. (2018). The role of mating context and fecundability in women's preferences for men's facial masculinity and beardedness. *Psychoneuroendocrinology*, 93, 90-102.

Feinberg, D. R., Jones, B. C., Smith, M. L., Moore, F. R., DeBruine, L. M., Cornwell, R. E., ... & Perrett, D. I. (2006). Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice. *Hormones and behavior*, *49*(2), 215-222.

Friedman, J., & Meares, R. A. (1978). Comparison of spontaneous and contraceptive menstrual cycles on a visual discrimination task. *Australian and New Zealand Journal of Psychiatry*, *12*(4), 233-239.

Frost, P. (1994). Preference for darker faces in photographs at different phases of the menstrual cycle: preliminary assessment of evidence for a hormonal relationship. *Perceptual and Motor Skills*, *79*(1), 507-514.

Gangestad, S. W., Haselton, M. G., Welling, L. L., Gildersleeve, K., Pillsworth, E. G., Burriss, R. P., ... & Puts, D. A. (2016). How valid are assessments of conception probability in ovulatory cycle research? Evaluations, recommendations, and theoretical implications. *Evolution and Human Behavior*, 37(2), 85-96.

Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014a). Do women's mate preferences change across the ovulatory cycle? A meta-analytic review. *Psychological Bulletin*, *140*(5), 1205.

Gildersleeve, K., Haselton, M. G., & Fales, M. R. cycle shifts in women's mate preferences: Reply to Wood and Carden (2014) and Harris, Pashler, and Mickes (2014). *Psychological Bulletin, 140*(5), 1272-1280.

Harris, C. R. (2011). Menstrual cycle and facial preferences reconsidered. *Sex Roles*, *64*(9-10), 669-681.

Harris, C. R. (2013). Shifts in masculinity preferences across the menstrual cycle: Still not there. *Sex Roles*, *69*(9-10), 507-515.

Harris, C. R., Pashler, H., & Mickes, L. (2014). Elastic analysis procedures: An incurable (but preventable) problem in the fertility effect literature. Comment on Gildersleeve, Haselton, and Fales (2014). *Psychological Bulletin*, 140, 1260–1264.

Havlicek, J., Roberts, S. C., & Flegr, J. (2005). Women's preference for dominant male odour: effects of menstrual cycle and relationship status. *Biology letters*, *1*(3), 256-259.

Holzleitner, I. J., & Perrett, D. I. (2017). Women's preferences for men's facial masculinity: trade-off accounts revisited. *Adaptive Human Behavior and Physiology*, 3(4), 304-320.

Izbicki, E. V., & Johnson, K. J. (2010). Dark, tall, and handsome: Evidence for a female increase in openness to other-race partners during periods of high conception risk. *Poster presented at the annual meeting of the Human Behavior and Evolution Society, Eugene, OR*. [Abstract only].

Johnston, V. S., Hagel, R., Franklin, M., Fink, B., & Grammer, K. (2001). Male facial attractiveness: Evidence for hormone-mediated adaptive design. *Evolution and human behavior*, *22*(4), 251-267.

Jones, B. C., Hahn, A. C., & DeBruine, L. M. (2018). Ovulation, sex hormones, and women's mating psychology. *Trends in Cognitive Sciences*.

Jones, B. C., Hahn, A. C., Fisher, C. I., Wang, H., Kandrik, M., Han, C., ... & O'Shea, K. J. (2018). No compelling evidence that preferences for facial masculinity track changes in women's hormonal status. *Psychological Science*, *29*(6), 996-1005.

Jones, B. C., Little, A. C., Boothroyd, L., DeBruine, L. M., Feinberg, D. R., Smith, M. L., ... & Perrett, D. I. (2005a). Commitment to relationships and preferences for femininity and apparent health in faces are strongest on days of the menstrual cycle when progesterone level is high. *Hormones and behavior*, *48*(3), 283-290.

Jones, B. C., Perrett, D. I., Little, A. C., Boothroyd, L., Cornwell, R. E., Feinberg, D. R., ... & Burt, D. M. (2005b). Menstrual cycle, pregnancy and oral contraceptive use alter attraction to apparent health in faces. Proceedings of the Royal Society B: Biological Sciences, 272(1561), 347-354.

Jünger, J., Kordsmeyer, T. L., Gerlach, T. M., & Penke, L. (2018). Fertile women evaluate male bodies as more attractive, regardless of masculinity. *Evolution and Human Behavior*, 39(4), 412-423.

Lewis, M. B. (2010). Why are mixed-race people perceived as more attractive?. *Perception*, *39*(1), 136-138.

Lewis, M. B. (2011). Who is the fairest of them all? Race, attractiveness and skin color sexual dimorphism. *Personality and Individual Differences*, *50*(2), 159-162.

Lewis, M. B. (2017). Fertility affects asymmetry detection not symmetry preference in assessments of 3D facial attractiveness. *Cognition*, *166*, 130-138.

Little, A. C., & Jones, B. C. (2012). Variation in facial masculinity and symmetry preferences across the menstrual cycle is moderated by relationship context. *Psychoneuroendocrinology*, *37*(7), 999-1008.

Little, A. C., Jones, B. C., & Burriss, R. P. (2007). Preferences for masculinity in male bodies change across the menstrual cycle. *Hormones and Behavior*, 51(5), 633-639.

Little, A. C., Jones, B. C., Burt, D. M., & Perrett, D. I. (2007). Preferences for symmetry in faces change across the menstrual cycle. *Biological psychology*, *76*(3), 209-216.

Little, A. C., Jones, B. C., & DeBruine, L. M. (2008). Preferences for variation in masculinity in real male faces change across the menstrual cycle: Women prefer more masculine faces when they are more fertile. *Personality and Individual Differences*, *45*(6), 478-482.

Lobmaier, J. S., Probst, F., Perrett, D. I., & Heinrichs, M. (2015). Menstrual cycle phase affects discrimination of infant cuteness. *Hormones and behavior*, *70*, 1-6.

Lobmaier, J. S., Sprengelmeyer, R., Wiffen, B., & Perrett, D. I. (2010). Female and male responses to cuteness, age and emotion in infant faces. *Evolution and Human Behavior*, 31(1), 16-21.

Marcinkowska, U. M., Galbarczyk, A., & Jasienska, G. (2018). La donna è mobile? Lack of cyclical shifts in facial symmetry, and facial and body masculinity preferences—A hormone based study. *Psychoneuroendocrinology*, 88, 47-53.

Marečková, K., Perrin, J. S., Nawaz Khan, I., Lawrence, C., Dickie, E., McQuiggan, D. A., ... & Imagen Consortium. (2012). Hormonal contraceptives, menstrual cycle and brain response to faces. *Social cognitive and affective neuroscience*, *9*(2), 191-200.

Navarrete, C. D., Fessler, D. M., Fleischman, D. S., & Geyer, J. (2009). Race bias tracks conception risk across the menstrual cycle. *Psychological Science*, *20*(6), 661-665.

Oinonen, K. A., & Mazmanian, D. (2007). Facial symmetry detection ability changes across the menstrual cycle. *Biological psychology*, *75*(2), 136-145.

Parlee, M. B. (1983). Menstrual rhythm in sensory processes: A review of fluctuations in vision, olfaction, audition, taste, and touch. *Psychological bulletin*, *93*(3), 539.

Pawlowski, B., & Jasienska, G. (2005). Women's preferences for sexual dimorphism in height depend on menstrual cycle phase and expected duration of relationship. *Biological Psychology*, *70*(1), 38-43.

Penton-Voak, I. S., & Perrett, D. I. (2000). Female preference for male faces changes cyclically: Further evidence. *Evolution and Human Behavior*, *21*(1), 39-48.

Penton-Voak, I. S., Perrett, D. I., Castles, D. L., Kobayashi, T., Burt, D. M., Murray, L. K., & Minamisawa, R. (1999). Menstrual cycle alters face preference. *Nature*, *399*(6738), 741.

Peters, M., Simmons, L. W., & Rhodes, G. (2009). Preferences across the menstrual cycle for masculinity and symmetry in photographs of male faces and bodies. *PloS one*, *4*(1), e4138.

Puts, D. A. (2005). Mating context and menstrual phase affect women's preferences for male voice pitch. *Evolution and Human Behavior*, *26*(5), 388-397.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301-308.

Sprengelmeyer, R., Lewis, J., Hahn, A., & Perrett, D. I. (2013). Aesthetic and incentive salience of cute infant faces: studies of observer sex, oral contraception and menstrual cycle. *PLoS One*, *8*(5), e65844.

Vaughn, J. E., Bradley, K. I., Byrd-Craven, J., & Kennison, S. M. (2010). The effect of mortality salience on women's judgments of male faces. *Evolutionary Psychology*, *8*(3), 147470491000800313.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632-638.

Van Wingen, G., Van Broekhoven, F., Verkes, R. J., Petersson, K. M., Bäckström, T., Buitelaar, J., & Fernández, G. (2007). How progesterone impairs memory for biologically salient stimuli in healthy young women. *Journal of Neuroscience*, *27*(42), 11416-11423.

Wood, W., Kressel, L., Joshi, P. D., & Louie, B. (2014). Meta-analysis of menstrual cycle effects on women's mate preferences. *Emotion Review*, *6*(3), 229-249.

Zietsch, B. P., Lee, A. J., Sherlock, J. M., & Jern, P. (2015). Variation in women's preferences regarding male facial masculinity is better explained by genetic differences than by previously identified context-dependent effects. *Psychological Science*, *26*(9), 1440-1448.

Table 1. Masculinity ratings for the generated faces. The scale ran from 1 (least masculine) to 9 (most masculine).

| | Mean Masculinity Score | Standard Error |
|---|---|---|
| Masculinised faces | 5.207 | 0.187 |
| Feminised faces | 4.528 | 0.157 |
| Africanised faces | 5.119 | 0.168 |
| Hyper-European faces | 4.607 | 0.179 |

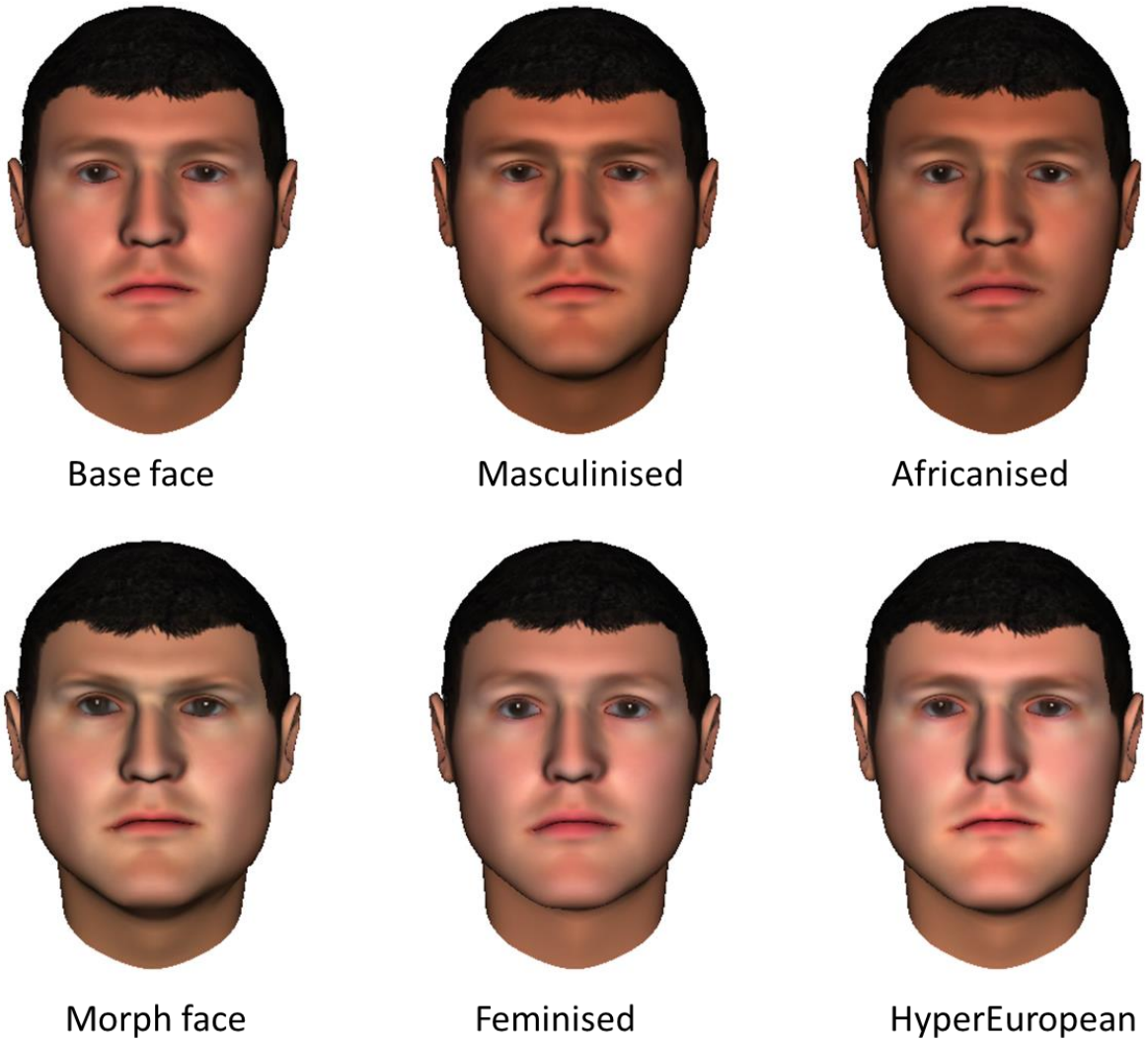


Figure 1. A set of faces used in the study.

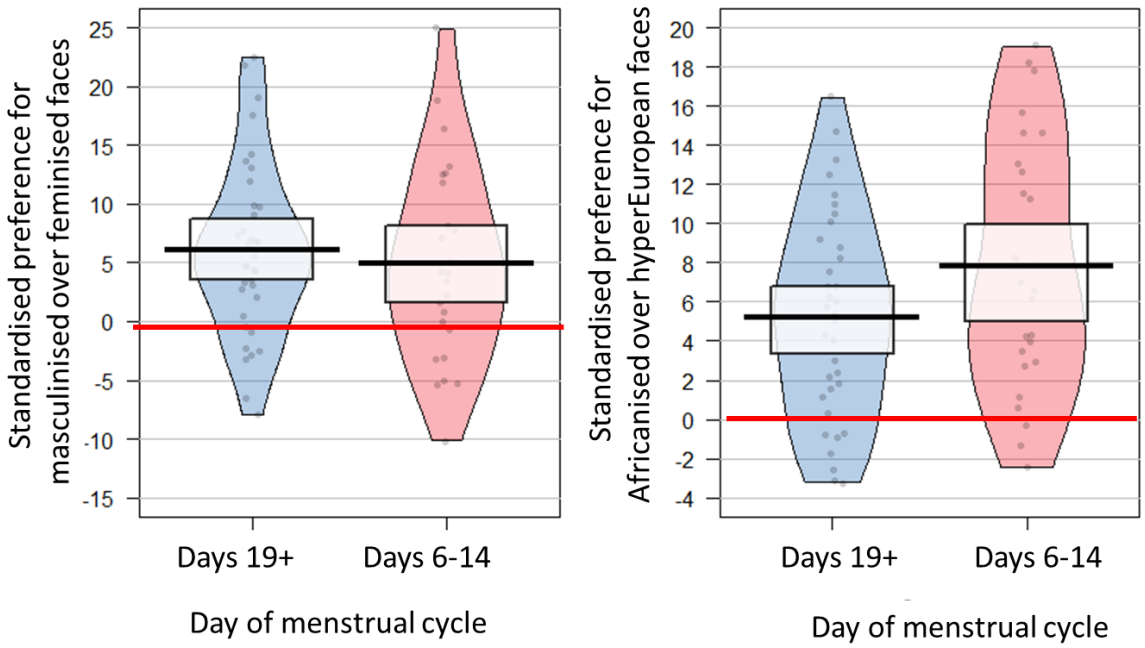


Figure 2. The size of the preferences for masculinised faces (left) and Africanised faces (right) as determined by attractiveness ratings in Task 1. The data are split according to whether the participants were in the fertile phase (days 6-14) or the non-fertile phase (days 19+). The red line indicates no preference and scores above this indicate a preference for masculinised faces or Africanised faces.

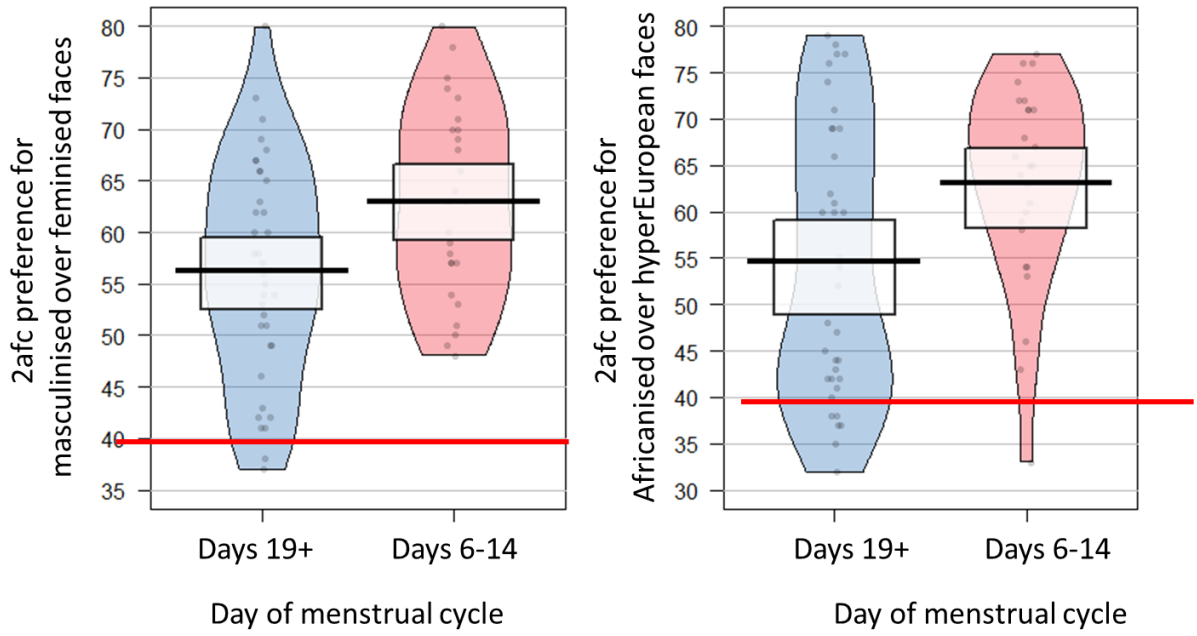


Figure 3. The size of the preferences for masculinised faces (left) and Africanised faces (right) as determined by 2afc responses in Task 2. The data are split according to whether the participants were in the fertile phase (days 6-14) or the non-fertile phase (days 19+). The red line indicates no preference and scores above this indicate a preference for masculinised faces or Africanised faces.

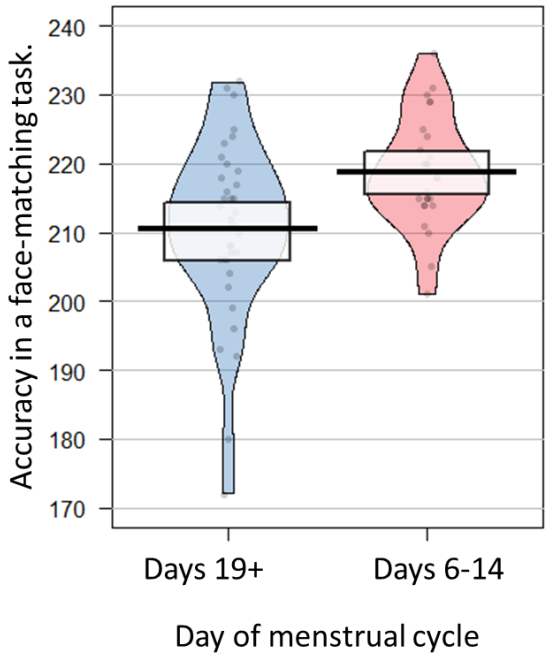


Figure 4. The accuracy scores for face matching performance in Task 3. The data are split according to whether the participants were in the fertile phase (days 6-14) or the non-fertile phase (days 19+). The maximum score was 240 and 120 was chance performance.

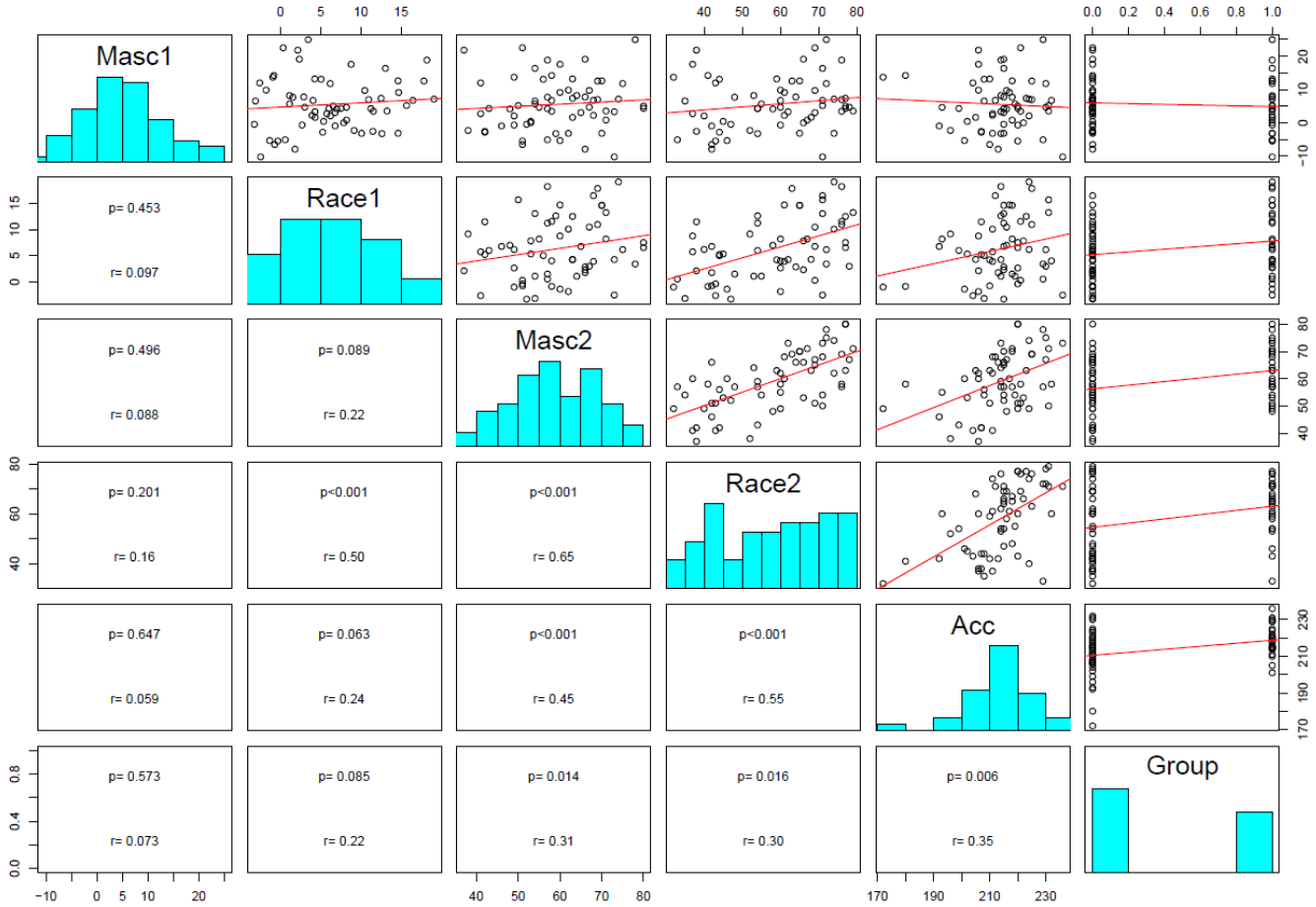


Figure 5. Correlations between the four preference measures, face-matching performance and the fertility group. *Masc1* is the masculinity preference based on attractiveness ratings. *Race1* is the Africanisation preference based on attractiveness ratings. *Masc2* is the masculinity preference based on 2afc selections. *Race2* is the Africanisation preference based on 2afc selections. *Acc* is the performance on the face matching task. *Group* refers to the fertility group for participants with the days 6-14 (fertile) group being coded as one and the days 19+ (non-fertile) group being coded as zero. The diagonal panels show histograms of the measures.

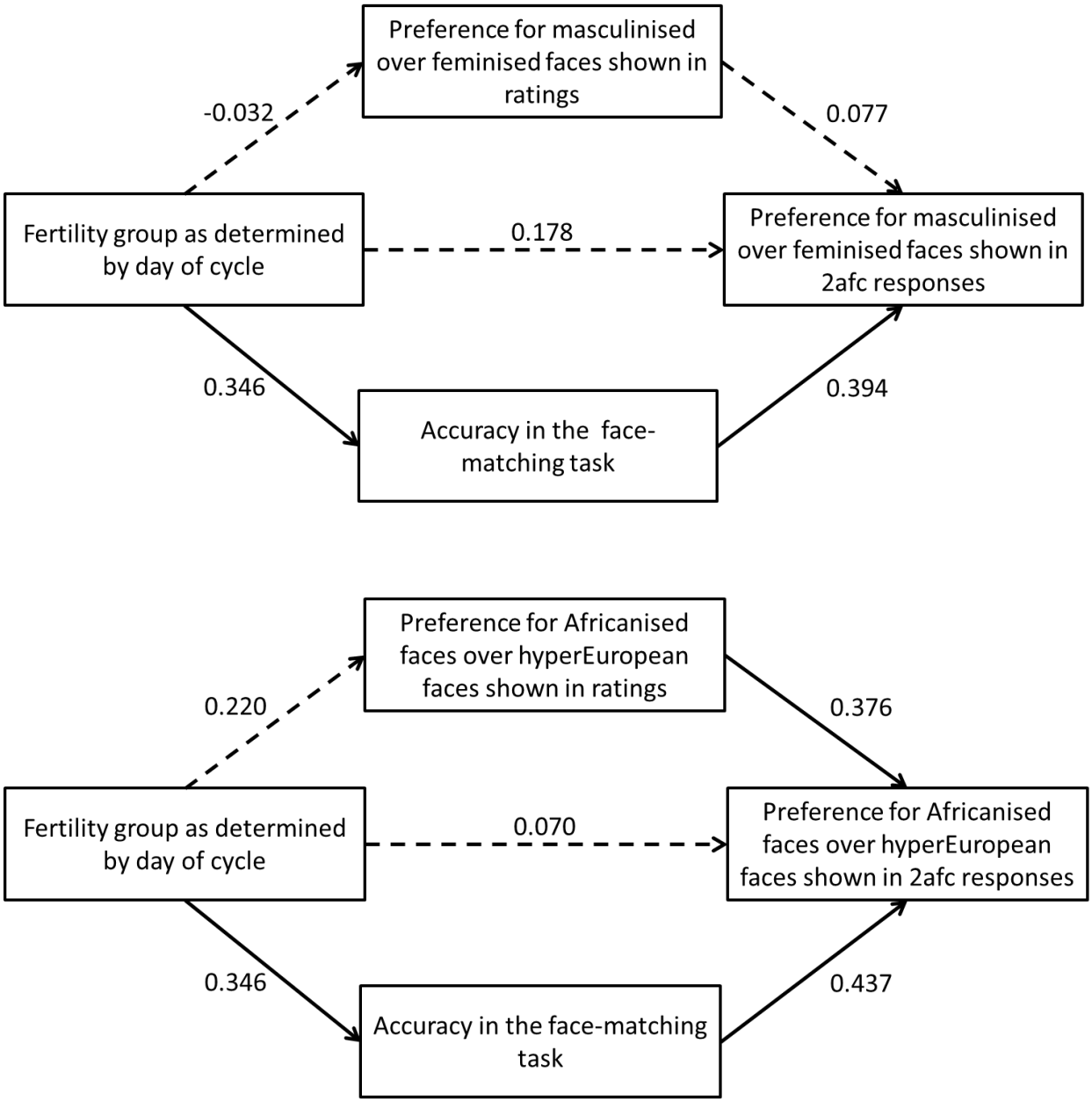


Figure 6. Path analyses showing the direct and mediated path from fertility to preferences based on 2afc designs for masculinised faces (Top) and Africanised faces (Bottom). The numbers show standardised beta coefficients. The solid arrows show significant paths whereas the dashed lines show non-significant paths. In both path analyses, the direct link between fertility and 2afc preference is mediated by face-matching accuracy.