This is a preprint of a manuscript that has been accepted for publication in Evolution and Human Behavior, doi: 10.1016/j.evolhumbehav.2019.08.005.

# No robust evidence for cycle shifts in preferences for men's bodies in a multiverse analysis: A response to Gangestad et al. (2019)

Julia Stern[1], Ruben C. Arslan[2], Tanja M. Gerlach[1], & Lars Penke[1]

[1]Department of Psychology & Leibniz ScienceCampus Primate Cognition
University of Goettingen
Gosslerstrasse 14, 37073 Goettingen, Germany

[2]Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany

Corresponding author: Julia Stern (julia.stern@psych.uni-goettingen.de)

## Abstract

Gangestad et al. (this issue) recently published alternative analyses of our open data to investigate whether women show ovulatory shifts in preferences for men's bodies. They argue that a significant three-way interaction between log-transformed hormones, a muscularity component, and women's relationship status provides evidence for the ovulatory shift hypothesis. Their conclusion is opposite to the one we previously reported (Jünger et al., 2018). Here, we provide evidence that Gangestad et al.'s differing conclusions are contaminated by overfitting, clarify reasons for deviating from our preregistration in some aspects, discuss the implications of data-dependent re-analysis, and report a multiverse analysis which provides evidence that their reported results are not robust. Further, we use the current debate to contrast the risk of prematurely concluding a null effect against the risk of shielding hypotheses from falsification. Finally, we discuss the benefits and challenges of open scientific practices, as contested by Gangestad et al., and conclude with implications for future studies.



Keywords: multiverse analysis; ovulatory cycle; mate preferences; steroid hormones; body masculinity; open science

Recently we[1] (Jünger, Kordsmeyer, Gerlach, & Penke, 2018) published a study in *Evolution and Human Behavior* showing that female preferences for cues of male body masculinity do not increase with fertility across the natural female ovulatory cycle, no matter if they are judged for attractiveness as a sexual or long-term partner. These results contradict the ovulatory shift hypothesis (Gangestad et al., 2005). Instead, we found some evidence for a general increase of female attraction around ovulation, independent of male body masculinity cues, which is in line with a general increase in sexual desire around ovulation (Arslan, Schilling, Gerlach, & Penke, in press) and the motivational priority shifts hypothesis (Roney, 2018). Gangestad and colleagues (this issue; henceforth Gangestad et al.) conducted a reanalysis on our open data, and although analyzing the same dataset, their results and conclusions differ significantly from ours. We appreciate Gangestad et al.'s effort and scrutiny of our data and analyses and welcome the opportunity to correct lapses in how we communicated our preregistered analysis. Still, we disagree that their reanalysis should lead to substantially different conclusions than the ones we stated. In the following, we clarify misrepresentations of our and Gangestad et al.,'s study and preregistration. Next, we provide a multiverse analysis, which provides evidence that Gangestad et al.'s results are not robust. We then discuss the risks of shielding a hypothesis from falsification and demonstrate the importance of open science practices.

## 1. Clarifying misrepresentations

Gangestad et al. critically address a number of points regarding our interpretation of our own preregistration, our analytic strategy and our conclusion. To begin with, Gangestad et al. criticize substantial parts of our preregistration. At the time we wrote our preregistration back in early 2016, preregistrations were not well-established in psychology and clear-cut

[1] Please note that we refer to the Jünger et al. (2018) results as "our results", although Ruben C. Arslan was not a co-author on this paper. Further, Julia Jünger's last name has since changed to Stern.

standards were lacking, especially for complex designs such as ours. As a consequence, we must admit that some parts of the preregistration were ambiguous and we agree that our preregistration left room for analytical flexibility. However, we disagree with their interpretation of our preregistration. We directly derived our analytical decisions from the wording of the hypotheses we preregistered. In the following, we will contrast our interpretation of our preregistration and our analytical decisions against those of Gangestad et al., criticise their analytical decisions that they claim to have derived from their preregistration, and clarify a potentially misleading reporting of an independent study by Marcinkowska and colleagues (2018b).

### *1.1. Predictor variables*

#### *1.1.1. Variables that might reflect body masculinity or muscularity*

In our study we investigated cycle shifts in preferences for seven potential cues of male body masculinity, including height, testosterone levels, strength, shoulder-chest ratio (SCR), shoulder-hip ratio (SHR), upper-torso volume relative to lower torso volume, and upper arm circumference. In additional analyses, we tested whether our effects were robust when controlling for BMI.

First, Gangestad et al. criticize our selection of variables and state that we did not offer a rationale for picking them. We are happy to expand on this. The stated aim of Jünger et al. (2018) was to clarify "whether there are mate preference shifts for masculine male body characteristics across the ovulatory cycle" (p. 413), thus conceptually replicating previous studies that reported ovulatory cycle shifts for preferences in body height (Pawlowski & Jasienka, 2005), sexual dimorphism in body shape (Little, Jones, & Burriss, 2007), and muscularity (Gangestad et al., 2007), especially in the light of reported null replications (Marcinkowska et al., 2018a; Peters et al., 2009). Note that Gangestad et al. deviate from our original article by moving the focus solely to muscularity. All seven male features we

preregistered and investigated were directly derived from previous evidence that they are sexually dimorphic in human adults and show links to formidability (e.g., Price et al., 2012). Detailed justifications including references can be found in the supplementary material (Table S1).

Second, Gangestad et al. point out that the simultaneous testing of all seven predictors in a multiple regression is a weak test for the potential effect of their shared variance, which undoubtedly exists. Yet we also analyzed a composite score variable, averaging all seven masculinity indicators, which did not change the results (see the open script on the Open Science Framework, https://osf.io/n4hj6/). Gangestad et al. ignored this additional analysis. Instead, Gangestad et al. compute a composite score of only two variables (strength and upper arm circumference), selected based on their associations with observer-rated bodily sexual attractiveness and dominance (the latter taken from the open data of Kordsmeyer et al., 2018[2]). Then they factor-analysed all variables and tested the hypotheses with one of the resulting factors as a robustness check. However, the composite score of strength and upper arm circumference, as used in the main analyses by Gangestad et al., includes only two out of seven preregistered masculinity predictors. Thus, we want to emphasize here that the lack of preference shifts for five out of seven body masculinity cues we preregistered seems uncontroversial and that Gangestad et al. shifted the focus to only two of them.

Third, Gangestad et al. claim that we did not properly control for confounding effects of BMI on preferences, because we controlled for a main effect of BMI, not an interaction effect. We agree that controlling for an interaction effect would have been the better way to control for confounds of preference shifts and thank Gangestad et al. for drawing attention to this issue. However, when we control for an interaction effect of BMI and cycle phase, the

[2] We would like to note that the bodily dominance ratings from Kordsmeyer et al. (2018) were collected after the Jünger et al. (2018) manuscript had already been submitted for publication, thus it never occurred to us to incorporate them into our original analyses, which would also have been a deviation from our preregistration.

estimated effects remain virtually identical and non-significant. Details can be found in the supplementary material (Table S2).

### *1.1.2. Cycle phase versus log-transformed hormones*

Further, Gangestad et al. criticize our sampling procedure and the decision to use cycle phase as our main predictor variable, as a number of fertile phase sessions might have been missclassified. Therefore, they claim that log-transformed hormone values would have been the better choice (section 4.12, Gangestad et al., this issue). First, cycle phase was clearly preregistered as our main predictor variable, as it was part of all of our hypotheses[3], whereas estradiol and progesterone were just mentioned in the mediator hypothesis. However, we used hormone levels for testing the mediation of our main effect, but not as mediators for the interaction effect, as we did not detect a significant interaction effect to be mediated (Baron & Kenny, 1986) and stopping the mediation test at this junction results in tighter error control. However, Gangestad et al. do not test a mediator effect either, as they simply regress the mediator on the outcome variable. Second, Gangestad et al. ignore our robustness analyses. More precisely, as a matter of fact, we excluded all of the potentially missampled participants, based on a combination of cycle regularity and LH test significance in our robustness checks. Thus, we redid all our analyses using this sample of $n$ = 112 women. Whereas it is true that a positive LH test alone does not necessarily indicate ovulation, using it together with a follow-up of the next menstrual onset[4] is probably one of the most reliable procedures we have to characterize the fertile phase (Fales, Gildersleeve, & Haselton, 2014; Gangestad et al., 2016).

[3] Just to give one example for a preregistered hypothesis tested in our study, the exact wording was "Moderation: When evaluating men as potential short-term partners based on their bodies, women in their fertile window, compared to their luteal phase, report increased attraction to men with higher baseline testosterone level". Hypotheses expecting an interaction effect were introduced with the word "moderation", hypotheses expecting a mediator effect of hormones were introduced with the word "mediation". The preregistration is publicly available at https://osf.io/egjwv/

[4] Also when ovulation is delayed and thus probably a second LH peak was undetected, the cycle must have been longer and characterized as irregular. Another reason for missclassification of cycle phase would be an anovulatory cycle, which would either lead to no positive LH test or, again, to a rather long, irregular cycle length.

In this subsample, the reported main effect of cycle phase became stronger, but the interaction effects that would be in favor of the ovulatory shift hypothesis still remained non-significant (Jünger et al., 2018, section 4.6), a fact that was not acknowledged by Gangestad et al.

Gangestad et al. claim that measures of salivary hormone levels are better predictors than a cycle phase variable comprised of LH tests and actual cycle length based on the reasonable assumption that estradiol and progesterone causally mediate the effects of cycle phase. However, they ignore the fact that we cannot measure salivary steroids with the same accuracy as LH surges. Crucially, measurement error can reverse which predictor is more likely to show an association. Indeed, since the liquid chromatography–mass spectrometry (LCMS) analysis of the estradiol levels only detected 22% of all possible values, the samples were reanalysed using an immunoassay kit (Jünger et al., 2018, p. 416). Interestingly, the correlation between LCMS analyses and the immunoassay data was $r = .06$, which made us doubt the reliability of the measures and underlined our preregistered decision to focus on cycle phase as a primary predictor. In line with this, Schultheiss, Dlugash and Mehta (2019) argue that estradiol and progesterone usually have extremely low concentrations in saliva, and are thus challenging to assess, even with LCMS analyses. They further mention that serum estradiol, when in a low range comparable to what is usually observed in saliva, can lead to immunoassay and LCMS outcomes that show unacceptably low convergence ($r = .32$, as reported in Huhtaniemi et al., 2012). Until recently the reliability of salivary hormone assessments might not have received much attention in the literature, but claiming that salivary hormones are *better* variables to investigate ovulatory cycle shifts compared to LH validated cycle phase with follow-up to the next menstrual onset requires ignoring the critical issue of measurement error. There is good evidence that LH tests can predict ovulation with high precision when compared to ultrasound-determined day of ovulation, which is usually

regarded as the gold standard (e.g. Blake, Dixson, O'Dean, & Denson, 2016), and much less evidence that salivary estradiol and progesterogen measures can do so.

Furthermore, even when deciding to use hormone values as a predictor rather than cycle phase, there are different ways to do so. Gangestad et al. decided to log-transform hormone values for certain theoretical reasons (which are debatable, e.g., Higham, 2016; Higham, this issue). In contrast, we simply centered hormone values within women and scaled them afterwards, which dealt with skewness (as shown in our Figure S1, and as previously done in other hormone-based cycle shift studies, e.g., by Jones et al., 2018; Roney & Simmons, 2016). A third possibility would be to use untransformed, raw hormone levels (as e.g., done by Marcinkowska et al., 2018a). All three approaches might have their advantages or disadvantages, so it is indeed difficult to decide what the best way is to deal with hormone values. Interestingly, when computing Gangestad et al.'s models using either scaled hormone values (as we did in our study) or untransformed hormone values instead of log-transformed values, the two-way interactions between E/P and their strength/muscularity component (S/M) as well as the three-way interactions between E/P, S/M and relationship status they report on become non-significant (all *p*s > .24; see Tables S3 and S4). Again, this fragility of their results was not acknowledged by Gangestad et alia.

### *1.1.3. Three-way interaction with relationship status*

Gangestad et al. criticize that we did not consider a three-way interaction effect with relationship status, although we preregistered it. It is correct that we did not report such an interaction. We decided not to report it as the simpler two-way interactions between cycle phase and masculinity cues (either entered individually, together, or as a composite score), were non-significant and test power was likely too low to detect a more complex three-way interaction effect (Mathieu, Aguinis, Culpepper & Chen, 2012; see also Section 1.2 below). We regret this omission as it is indeed a deviation from our preregistration, but saw it as

permissible at the time because it led to unaltered conclusions (see Table S5). Even in Gangestad et al.'s reanalysis, the two-way interactions between S/M and ln(E/P), or S/M and ln(E) or ln(P), printed bold in their Tables 4, 5 and 6, because they are "primary effects of interest", are almost all non-significant. Importantly, the majority of effects even point in a negative direction, opposite of the expected effect. Additionally, Gangestad et al.'s analyses of the three-way interaction of cycle phase x S/M x relationship status do not result in a significant effect (see their Table 9). The three-way interaction effect they focus on is a different one: "We include the ln(E/P) x Strength/Muscularity x Relationship Status interaction. This hypothesis had been specified in Jünger et al.'s pre-registration but was not tested in their analysis" (Gangestad et al., p. 6). This is not true: we preregistered a three-way interaction involving cycle phase, relationship status and the masculine body cues. Neither a Strength/Muscularity composite or factor, nor the three-way interaction involving hormones, nor the log-transformation of E/P was part of our preregistration. Gangestad et al. make it seem as if we file-drawered results that ran counter to our favored conclusion, but we never preregistered, nor ran any of the analyses that yielded significant findings in Gangestad et al. (i.e., mainly the three-way interaction between ln(E/P), S/M and relationship status, controlling for BMI, on sexual attractiveness ratings).

In addition, we also disagree that their reported analysis on the effect of the three-way interaction between ln(E/P), S/M and relationship status, controlling for BMI, on sexual attractiveness ratings maps onto the theoretical predictions we made in our paper. In our preregistration, we predicted that cycle shifts in preferences are larger for partnered women than for single women (Hypothesis 7, p. 6). A simple p-value for a three-way interaction does not answer this question; the interaction has to be unpacked. When doing so by analyzing the two-way interactions between log-transformed hormones and the muscularity composite score, Gangestad et al. report that the effect is positive but non-significant for partnered

women, whereas it is negative and significant for singles (see their Table 6). Both effects have the same size of an unstandardized model estimate (0.03 on an 11-point Likert scale), but in opposite directions. Based on the theory, we would expect a strong interaction in partnered women, and an attenuated or zero interaction in single women, not the cross-over effect reported by Gangestad et al. (as Gangestad et al. acknowledge).

Furthermore, even for Gangestad et al.'s preferred main result the effect size is not very impressive. Gangestad et al.'s Figure 1 shows model-based estimates of the associations at the 5th and 95th percentile of S/M. Even when choosing such extreme values for the moderator, the interaction is barely apparent in their graph. Below, we show a slightly different graph (see Figure 1) of the same model in which we display model-based estimates of the effect of the S/M component by relationship status and average versus high log(E/P). We superimpose (in gray) the model-based differences between women in the strength of the association (random slopes). We think this graph supports our view that there is only little variation between and within women in the preference for S/M. Even using Gangestad et al.'s preferred model, it seems clear that the purported moderators (ln(E/P) and relationship status) explain little of this variation between and within women. Although Gangestad et al. are correct in saying that our reported Spearman rank correlation does not preclude cycle changes in preferences, we think the graph rather supports our interpretation.

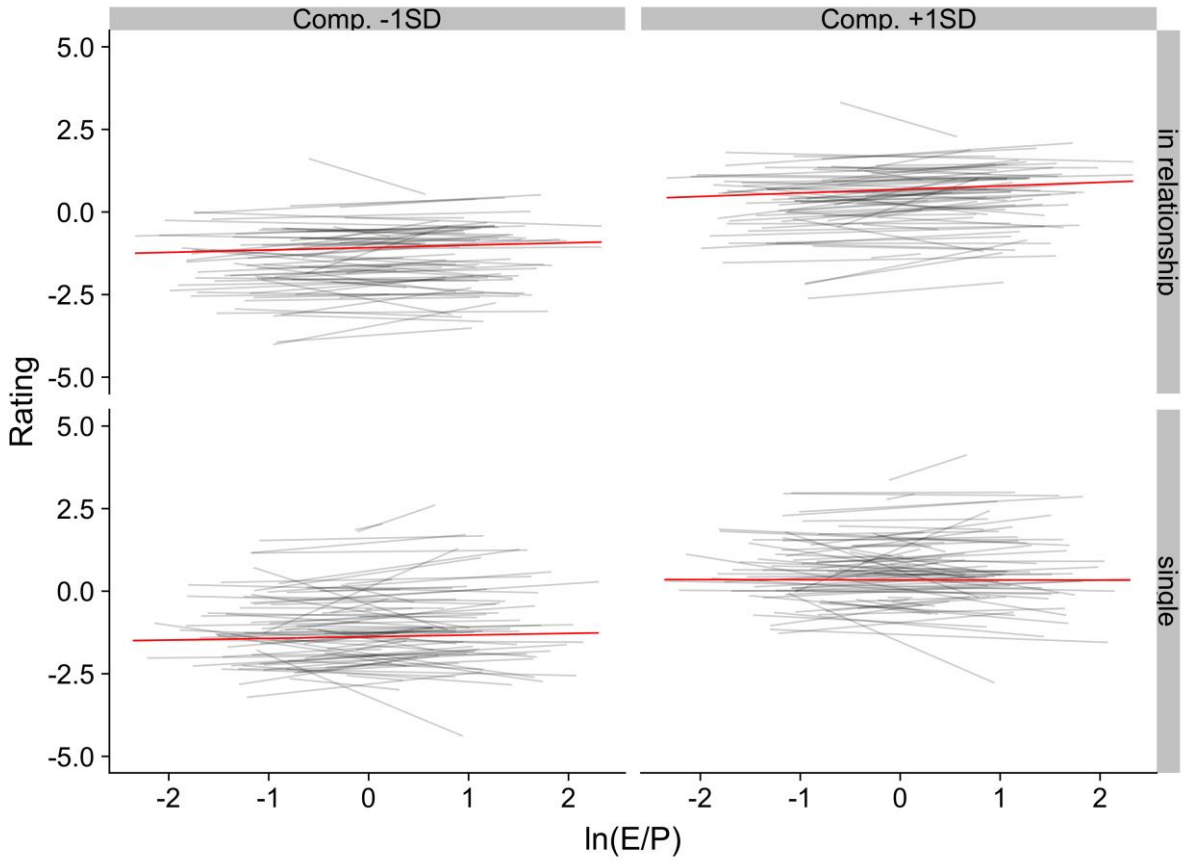


**Figure 1.** This spaghetti plot shows that only a very small amount of the variation in slopes between women (gray lines) is explained by the moderators ln E/P and relationship status. For the most part, women consistently prefer men who are higher in muscularity (Gangestad et al.'s S/M component). The slopes are extracted from the fitted multilevel model from Gangestad et al.'s Table 3 and are estimated adjusted for BMI. The mean levels in this marginal effect plot reflect an average BMI man.

### *1.2. Gangestad et al.'s preregistration*

Gangestad et al. want to show that their analyses are not data-dependent and thus comparable in informational value to our preregistered analyses. To substantiate this, they base some of the analytic decisions they aply to our data on a preregistration for a separate, but somewhat similar study of theirs that they uploaded to the Open Science Framework on 18 April 2018 (https://osf.io/4x7ub/). This is important, because it could potentially ensure that their analytic decisions were not biased by seeing our results. However, clearly the decision to re-analyse our data at all was made after seeing our study and our results, as was the decision to frame

the re-analysis in terms of parts of their own preregistration. The impact of such a case of potential partial data-dependence is hard to predict and it is not clear how well overfitting is still guarded against (see also Jones, Marcinkowska & DeBruine, this issue). More importantly, the way they modelled the three-way interaction of log-transformed E/P x muscularity component x relationship status, controlling for BMI, on sexual attractiveness ratings, which is the main analyses they built their reanalysis on, is actually not even part of Gangestad et al.'s preregistration for their separate study, as their study is based on morphed stimuli for which a Strength/Muscularity component or factor cannot be computed, nor was a BMI control necessary or planned for their morphed stimuli. Furthermore, in their preregistration, they explicitly describe a two-way interaction as their key hypothesis, as they aim to primarily recruit women in relationships, not singles. Thus, contrary to their claim, the exact analyses they did were never preregistered by anyone.

Moreover, we want to draw attention to the fact that in their preregistration, Gangestad et al. provide a power simulation, which is laudable. This power simulation indicates that, with $N = 250$ women, they have a test power of .94 to detect a two-way interaction effect of $d = .35$. Transferred to the analyses they report in their reanalyses ($N = 157$ women, a three-way interaction effect and a much smaller effect size), their analyses seems heavily underpowered to find the effect they are reporting. This increases the risk that effect sizes are overestimated, thus making their reproducibility questionable (e.g., Button et al., 2013). At the very least, the three-way interaction they report requires direct replication in a well-powered study before any weight can be put on it.

In summary, Gangestad et al. refer to their own preregistration to lend credence to the idea that their re-analysis of our data was just as unbiased by seeing the data as were ours. This is misleading, because important analytic decisions, crucial for the pattern they report, were made after seeing our results and data. At best, a subset of decisions was constrained by

their preregistration. As it stands, their analyses and reporting gave Gangestad et al. much leeway to pick and choose which *p*-values to focus on. Combined with the lower power to detect realistic effect sizes for moderators according to their own power analysis, their results are probably not robust.

### *1.3. Gangestad et al.'s "independent demonstration": misrepresenting Marcinkowska et al.'s (2018b) results*

In section 5.7 of their reanalysis, Gangestad et al. report an effect of Marcinkowska et al.'s (2018b) study. Here, they state that Marcinkowska et al. (2018b) report a similar three-way interaction as they find, claiming that "these results give additional reason to think that the interaction effect we report is robust" (p. 14). Note that this is the same dataset in which Marcinkowska and colleagues did not observe any compelling evidence for any hormonally influenced within-woman preference shifts across the cycle for facial masculinity, facial symmetry or body masculinity (reported in a different article, Marcinkowska et al., 2018a). Marcinkowska et al. (2018b) mainly focus on between-women effects, but also report a number of different robustness checks for within-women hormone effects, all finding no compelling evidence for preference shifts across the cycle or tracking changes in within-woman hormone levels. There is one exception. In Table S24 (in their supplementary material) they report a significant interaction effect between daily progesterone levels and relationship status on preferences for masculine bodies ($p = .04$). They further report that simple effect analyses suggest that this effect is positive and only significant for singles ($p = .01$), not for paired participants ($p = .96$). Note that this effect thus runs in the exact opposite direction of the effect Gangestad et al. report for our dataset. Thus, the one singled-out significant result from Marcinkowska et al.'s (2018b) extensive supplementary robustness checks (31 Tables) does not support the robustness of the three-way-interaction Gangestad et al. found in our data.

## 2. Using multiverse analysis to increase transparency

Above, we hinted that changing almost any single analytical decison in Gangestad et al.'s analysis leads to non-significant results. But which analytical decisions are the right ones? That is probably impossible to tell, because many potential decisions are plausible and several may even be equally right in the sense that they provide approximations of the construct of interest. The concept of the garden of forking paths (Gelman & Loken, 2013) explains how researcher's decisions can lead to a multiple comparisons problem via considering a large number of potentially plausible analytical decisions. Thus, it explains how our results can differ from those reported by Gangestad et al. despite analyzing the exact same data. In their Table 2, they describe the key differences between their and our analytical choices. Here, we take the opportunity to translate these differences to possible and plausible decisions that have to be made when walking through the garden of forking paths. The directly derived choices from these differences are displayed in Table 1.

Table 1

Differences in Gangestad et al.'s and our analytical choices that lead to different paths in the garden of forking paths

1. Predictor 1: Assessment of fertility
   a) Cycle phase whole dataset ($N$ = 157)
   b) Cycle phase LH validated dataset ($n$ = 112)
   c) Hormone levels: log-transformed hormones
   d) Hormone levels: mean-centered and scaled hormones
   e) Hormone levels: raw hormone levels
2. If fertility assessed by hormone levels, how are they entered?
   a) Estradiol-to-progesterone ratio
   b) Estradiol and progesterone separately
3. Predictor 2: masculinity/ muscularity cue
   a) Factor analysis, resulting in 3 factors
   b) „Empicirally vetted" strength / upper arm circumference composite
   c) Simultaneous entry of all 7 variables
   d) Composite score of all 7 variables
   e) Separate models for all 7 variables
4. Control variable
   a) Controlling for an interaction effect of BMI
   b) Not adding a control variable
5. Two-way vs. Three-way interaction

a) Three-way interaction with relationship status
b) Two-way interaction without relationship status

Figure 2 shows a garden of forking paths: it showcases the possible analytical decisions regarding our dataset that are displayed in Table 1. Please note that this graph only shows possible plausible decisions after already deciding for cycle phase as a predictor variable, which are approximately 1/4$^{th}$ of plausible analytical decisions we focus on here. The reason we did not display the decision for hormone variables here is that the figure involving all decisions was simply too big to be printed (and would require at least A2 format). The full garden of forking paths can be found in the supplementary material (Figure S1).

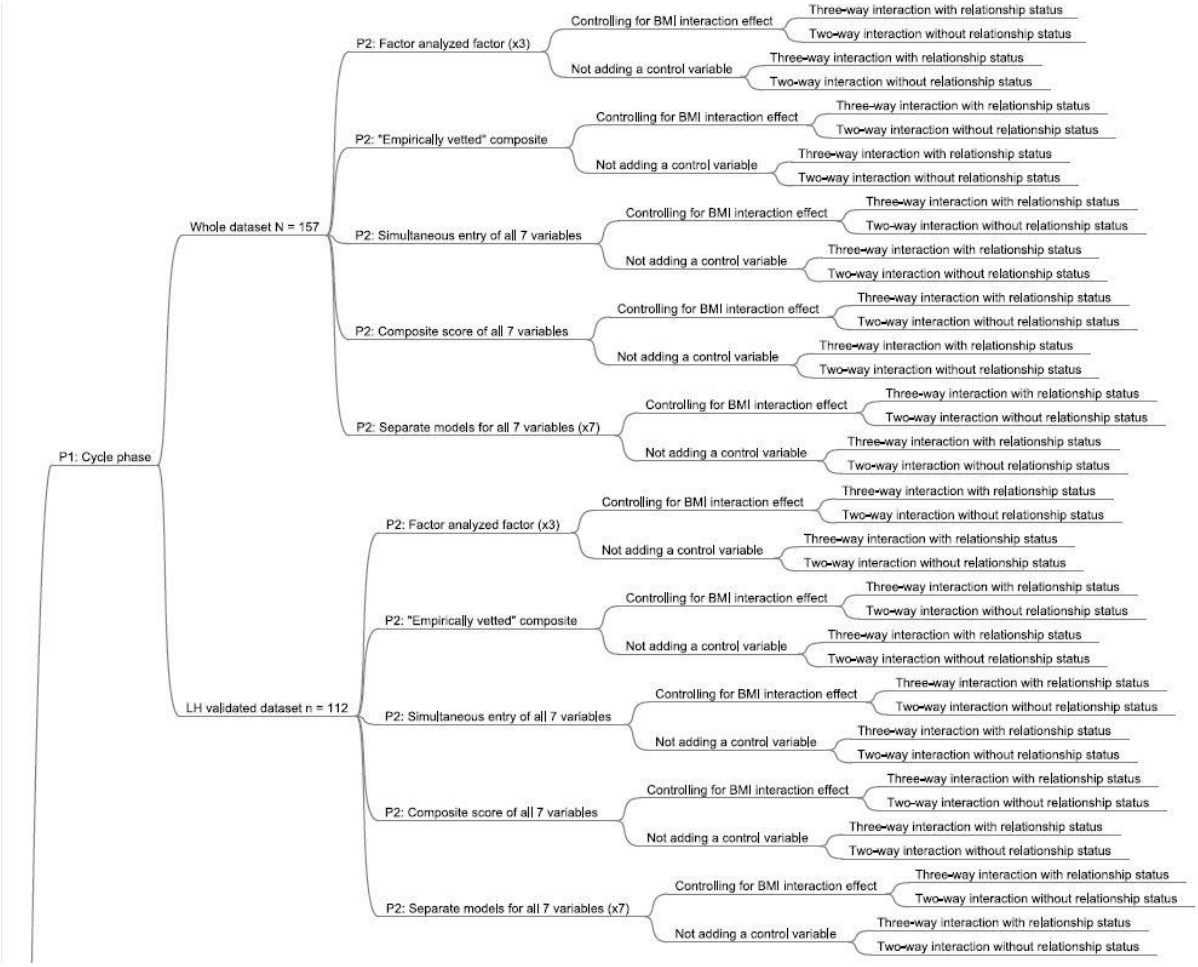


**Figure 2.** A graphical representation of a garden of forking paths, illustrating possible and plausible analytical decisions after deciding for cycle phase as a predictor for cycle shifts in

preferences. Note that this figure only displays approximately 1/4th of the possible and plausible decisions. The full garden of forking paths can be found in the supplementary material (Figure S1).

Our preregistration did not specify statistical models. This can be seen as allowing ourselves many researcher degrees of freedom, making it easier to reveal foregone conclusions. Of course, we believe we tested models that were reasonably based on the literature and did not try to engineer a particular conclusion. Moreover, we had several robustness checks in our paper (e.g., repeating the analyses with $n$ = 112 women with LH validated fertile phase, using separate models for all cues, and generating a composite score averaging all cues), thus already protecting against arbitrary analytical decisions, more so than is usually done in the literature. However, our private beliefs and internal best practices can hardly stand up to the level of scrutiny in Gangestad et al.'s critical commentary. Therefore, we decided to run a multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016) to investigate whether the null results for preference shifts we previously reported (Jünger et al., 2018) or Gangestad et al.'s reported effects are more robust (or whether neither are). A multiverse analysis entails making all the different analytical decisions that would be possible and plausible for a given hypothesis and then running all the respective statistical tests (Steegen et al., 2016). The resulting *p*-values of all these analyses are then displayed in a single histogram. More precisely, we investigate whether choosing a different path during the data transformation or analytical decision process has a significant impact on the results and how many of the different analyses do, indeed, lead to statistically significant results. Thus the resulting large set of reasonable scenarios will show how conclusions can change because of arbitrary analytical decisions.

How do we construct such a multiverse of decisions? After all, there already are almost infinite possible decisions about what counts as an outlier to exclude. To construct this

multiverse in a principled manner, we focused on the decisions where we and Gangestad at al. took different turns in the garden of forking paths that were reported as "primary" differences in their Table 2. This does not, by any means, exhaust all plausible possibilities. One could easily argue that, for example, including or excluding between-women hormone effects, other control variables (such as testosterone levels), different random slope specifications, and so on might be additional plausible decisions. However, all the different decisions that they refer to, shown in Table 1 and Figure S1, already led to 416 different models and 1,254 *p*-values of interest[5]. We computed all these different models. Data and analysis script for the multiverse analysis is publicly available (https://osf.io/6afhg/).

As displayed in Figure 3, the results suggest that any cycle shifts in mate preferences for men's bodies reported in Gangestad et al. might not be robust: Out of 1,254 resulting *p*-values, 31 were significant (<.05), thus 2.47 percent. One could think that these significant *p*-values all stem from small variations of the model Gangestad et al. report and do, thus, indicate robustness of their results. This is not the case. Rather, they stem from very different paths and about half of them even point in the direction opposite of what is predicted by the ovulatory shift hypothesis. Details can be found in Table S6.

Further, we want to stress that *p*-values, by their nature, are distributed equally (as they are equally likely) when the null hypothesis is true. If an effect exists, the distribution of significant *p*-values should be right-skewed, even when the effect is small and test power to detect it is low (Simonsohn, Nelson, & Simmons, 2014). However, the rate of 2.47% significant *p*-values from our analysis is even below the rate of 5% significant *p*-values one would expect by chance as false positives. Furthermore, the overall distribution is rather uniform, whereas the significant *p*-values <.05 are left-skewed, not right-skewed as would be

[5] Note that in most models more than one *p*-value is of interest: Models with E and P separately entered have at least two, models with the seven predictors entered simultaneously have at least seven and models testing a three-way interaction also contain a *p*-value for a two-way interaction.

expected for a robust effect. Note that the effect Gangestad et al. report in their main analysis ($p$ = .019, see their Table 4) is the smallest $p$-value in our multiverse analysis (see Table S6). How come the effect Gangestad et al. reported is framed as robust by them? Indeed, most of the models they report are miniscule deviations from their analytical decisions (e.g. including third variables such as testosterone or age as controls, which neither we nor they ever discussed as central), but do not really reflect a difference in the primary analytical decisions as displayed in their Table 2, which we combined in our multiverse analysis.

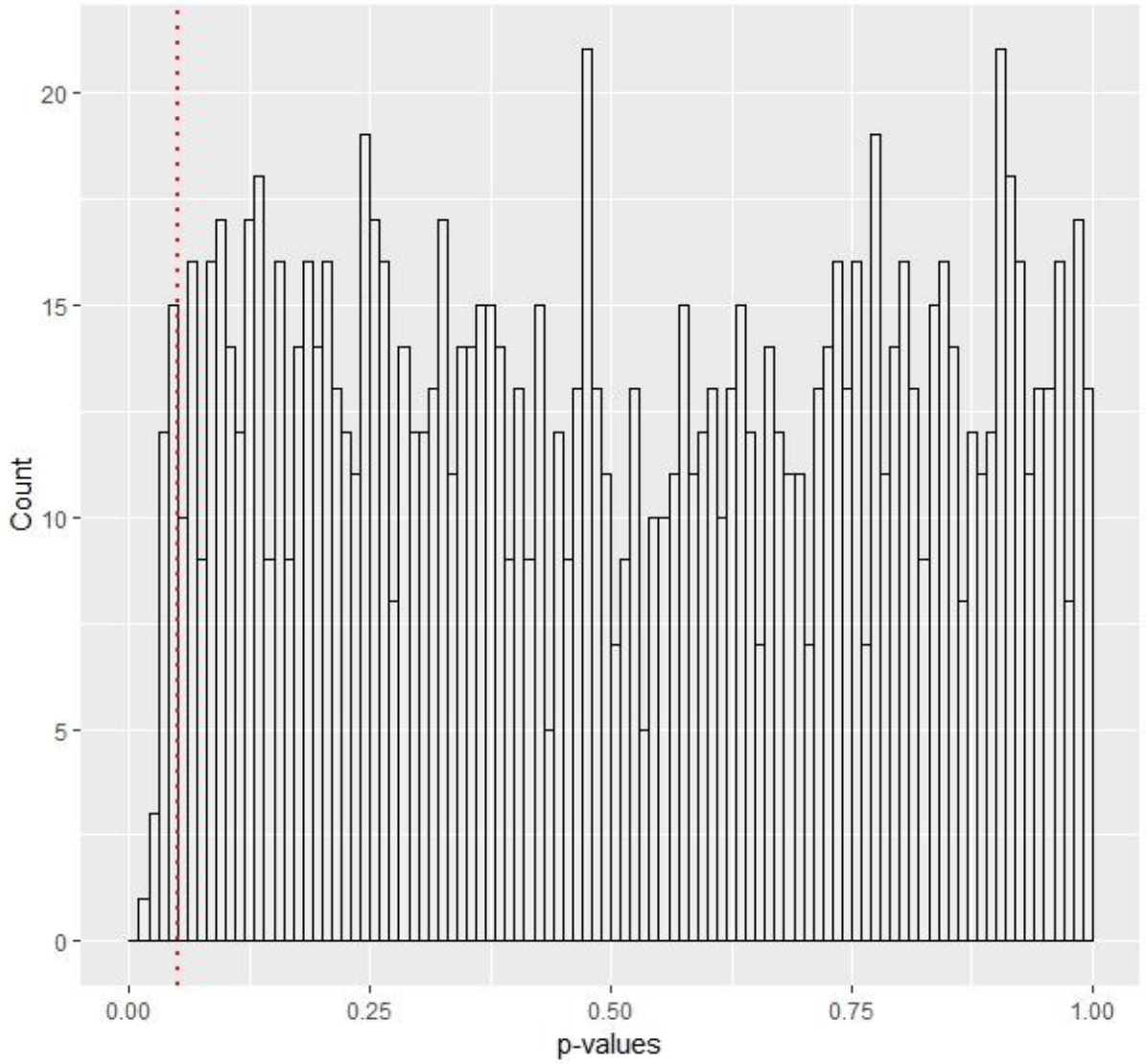


**Figure 3.** Histogram displaying the frequency of the 1,254 p-values of interest resulting from the multiverse analysis. Note that the red dotted line is at $p$ = .05 and thus separates nominally significant results on the left from nominally non-significant results on the right.

## 3. The problem of unfalsifiability

The good genes ovulatory shift hypothesis (proposed by Gangestad et al., 2005) has been tested in quite a number of studies (meta-analysed in Gildersleeve, Haselton & Fales, 2014, and Wood et al., 2014). As stated in Gildersleeve, Haselton and Fales (2014), the ovulatory shift hypothesis makes three directly testable predictions: First, when fertile, women should be more sexually attracted to men's characteristics that reflect good genes, compared to their low-fertility days. Second, cycle shifts in women's mate preferences for good genes characteristics should be absent or only weakly present when evaluating men for long-term relationships. Third, when fertile, women should not be sexually attracted to men's characteristics that reflect a higher suitability as a long-term partner, compared to their low-fertility days.

Since it is not possible to test the third prediction here (as there is no clear hypothesis regarding which characteristics in men's bodies should reflect a higher suitability as a long-term partner), we will focus on the other two predictions. Regarding the first prediction, we did not find compelling evidence that women's mate preferences vary across the cycle (or on high-fertility compared to low-fertility days). Women's cycle phase did not, neither in our original study, nor in Gangestad et al.'s reanalysis, nor in our multiverse analysis, interact significantly with any of the assumed indicators of good genes (i.e., cues of body masculinity/muscularity) to predict sexual attractiveness ratings. When choosing hormones as a predictor variable rather than cycle phase, the two-way interaction between hormone levels and the purported indicators of good genes were also non-significant. However, Gangestad et al. reported a significant three-way interaction with women's relationship status. Importantly, this interaction effect was only significant when log-transforming hormone levels and in combination with other analytical decisions, e.g., computing a certain composite score and controlling for BMI. When unpacking this three-way interaction, Gangestad et al. report that

the effect was only significant for singles, not for partnered women, and in the opposite direction as predicted by the ovulatory shift hypothesis (though it was in the predicted direction for partnered women). Still, our multiverse analysis suggests the effects reported by Gangestad et al. are not robust.

Regarding the second prediction, our and Gangestad et al.'s results point in the same direction: results for long-term attractiveness do not differ from results for sexual attractiveness. Indeed, the effect is absent when evaluating cycle phase as a predictor of long-term attractiveness, but given that the same is true for sexual attractiveness, this result cannot be seen as in favor of the ovulatory shift hypothesis. Moreover, for those log-transformed hormone analyses for which Gangestad et al. found significant effects for sexual attractiveness, the same effects were significant for long-term attractiveness ratings (see their Table S20). They fail to mention this. This raises the question of how their results can be in favor of their hypothesis, if results for sexual and long-term attractiveness are virtually identical. However, Gangestad et al. might argue that there are no long-term attractiveness cues in bodies that are independent from sexual attractivess cues.

Let us evaluate the evidence. Gangestad et al. seem to agree with us that there are no ovulatory preference shifts on individual cues to body masculinity or sexual dimorphism, such as height, contradicting some earlier studies (Little, Jones, & Burriss, 2007; Pawlowski & Jasienka, 2005). When the focus is shifted to upper-body muscularity, we begin to disagree. In our analyses we find no evidence for preference shifts at all. Gangestad et al. find significant effects for a set of analyses with very specific assumptions about how to construct the muscularity variable, what to control for, how to conceptualize ovulation (on a very proximate level), how to transform variables, and how to specify the multilevel model. Contrary to their claims, most of these analytic decisions are not constrained by either their or our preregistration. Gangestad et al. give extensive justifications for each of their analytic

decisions, but our multiverse analysis makes it clear that virtually all other reasonable sets of analytic decisions do not lead to significant results. Of course it might be the case that Gangestad et al. have indeed identified the most ideal set of analytic decisions, but then it is still peculiar that their significant effect is so fragile that it immediately breaks down under most reasonable variations of the analytic decision, especially given that our data provide more statistical power than most previous studies. For these reasons, we do not think that our data and results, nor the results reported by Gangestad et al., are in favor of the ovulatory shift hypothesis. Indeed, the null results of our study are in line with other, recently published, large-scale replication studies investigating cycle shifts in preferences for masculine faces (Dixson et al., 2018; Jones et al., 2018; Marcinkowska et al., 2018a), bodies (Marcinkowska et al., 2018a; van Stein et al., 2019), voices (Jünger et al., 2018b) and behaviors (Stern, Gerlach, & Penke, 2019). Drawing null conclusions from just our data would be premature. However, recent work clearly challenges previous evidence for the ovulatory shift hypothesis, especially because recent studies used more rigorous methods and designs than previous reports of significant effects (for an overview see Jones, Hahn, & DeBruine, 2019). This clearly shifts the balance to a need for more positive evidence in order to retain the good genes ovulatory shift hypothesis.

But even if the three-way interaction between hormones, upper-body muscularity and relationship status on sexual attractiveness ratings was robust, that does not imply that it is practically meaningful. We agree with Gangestad et al. that just focussing on p-values and setting a rather arbitrary cut-off (e.g., $p < .05$) to decide about the existence of an effect (what they call “simple up-down thinking”, p. 14) is problematic for several reasons already outlined by Gangestad et alia. We agree that it is also important to include effect sizes. Thus, we encouraged Gangestad et al. during the review process to specify the smallest effect size of interest (SESOI; Anvari & Lakens, 2019 ) that would still be consistent with an adaptive

evolutionary explanation, and hence in favor of the hypothesis. In section 5.4. Gangestad et al. state that "the current data do not allow one to pinpoint effect sizes with sufficient precision to judge their theoretical meaningfulness or practical impact" (p. 13). The reported unstandardized effect size of their three-way interaction was 0.05 on an eleven-point Likert scale. Although we agree that "headless digital figures" (p. 34) might not have the same effect as real-life male bodies, this statement, together with the previously raised issues, shields their hypothesis from falsification. If we cannot falsify the hypothesis based on p-values or effect sizes, or the overall evidence provided by recent, rigorous studies, how could we ever do so? If it is not possible to falsify a hypothesis, is it even possible to confirm it?

We agree with Gangestad et al. that null conclusions can discourage future research on a topic. We agree that one should not make strong conclusions in favor of the null hypothesis too early, especially not based on a single study. We agree that more data is needed from independent, highly powered, preferably preregistered, replication studies employing strong methods and designs. Regarding the current evidence, we are happy to conclude uncertainty about the effect. However, it should be noted that most of the original significant findings in the earlier literature come from underpowered studies, making them at least in need of replication. All recent high-powered replication studies did *not* find compelling evidence for the effect. Statistical tests of more complex hypotheses, like the moderation by relationship status, were probably underpowered in all existing studies so far. Hence, we encourage researchers to collect more data on this research question. However, we also urge researchers to specify testable, falsifiable hypotheses and standards for falsification, as unfalsifiable hypotheses impede scientific progress, the search for alternative hypotheses, and thus the accumulation of knowledge.

**4. Showcase for the importance and helpfulness of Open Science**

Gangestad et al. are concerned that studies using open scientific practices might be prematurely evaluated positively without appropriate scrutiny (p. 37). While we take this concern seriously, we also think the current exchange clearly demonstrates the advantages of open science, as it would have not been at all possible without embracing open science practices. The more researchers publicly offer about the planning and hypothesis of a study (in the form of a preregistration or registered report), the data, analytic code, and material, the better the study can be critically checked and independently evaluated. This can also motivate researchers to increase the quality of their work. We agree with Gangestad et al. that preregistration does not ensure appropriate testing of hypotheses or meaningful results. It certainly is also not in itself a guarantee for well-conducted research or high data quality. Most preregistrations are, indeed, improvable, including ours for the current study. We clearly learned over the last few years that writing a good, precise preregistration is hard, especially for complex research designs and hypotheses. Still every little bit of added transparency helps, as every bit reduces researcher degrees of freedom. In garden of forking path situations, the main thing we want to avoid is choosing the path based on the outcome, i.e., whether a hypothesis is supported or falsified. Therefore, preregistration prevents a number of questionable research practices. In addition, we think that review before results, as in the increasingly popular format of Registered Reports (Chambers, 2013), can clearly improve scientific practice. Importantly, as many authors in the open science literature have pointed out, this does not negate the value of exploratory research. Exploration is often useful and necessary, but to avoid misleading ourselves, strategies to prevent overfitting, including replication, controlling for multiple testing, or dividing the data into training and test sets are very important. Further, transparency is crucial: exploratory analyses should be framed as exploratory. Reporting selected *p*-values from exploratory research, on the other hand, has more potential to mislead than to enlighten.

This valuable post-publication discussion of our work sheds light on many underdiscussed decisions in data analysis and scientific practice. Although we ultimately disagree that Gangestad et al.'s re-evaluation of our work leads to substantially different conclusions, we are glad that open data and preregistration enabled this discussion. Importantly, many of the researchers of recently published studies investigating ovulatory cycle shifts (Dixson et al., 2018; Jones et al., 2018; Jünger et al., 2018a; 2018b; Stern et al., 2019) opened their data, allowing for in-depth evaluations of the conducted analyses and the conclusions put forward, as shown in the current debate. However, all studies for which open data were provided reported no compelling evidence for the ovulatory shift hypothesis. In sharp contrast, none of the studies reporting evidence in favor of the hypothesis opened their data, making it impossible to evaluate whether any previously reported evidence is, indeed, robust. Hence, we not only encourage authors of future studies, but also of previous studies to open their data and analytic scripts, as we think this is the only way to fairly evaluate the whole picture. We need to subject the literature that provided support for the effects on which this discussion is based to the same level of scrutiny applied here to make progress. We agree that open science practices alone are not an indicator of research quality, but all else being equal, a more transparent study has a higher potential to make a lasting contribution to our knowledge.

We are happy that our study shows both the benefits and the challenges of open science. We think that this process clearly demonstrates the importance of transparency and we hope that it helps to make future science more open and reproducible.

**Data availability**

Open data, open analysis script and the supplementary material are publicly available at https://osf.io/6afhg/

## References


Anvari, F., & Lakens, D. (2019). Using anchor-based methods to determine the smallest effect size of interest. *Preprint on PsyArXiv.* Retrieved from https://psyarxiv.com/syp5a/ Doi: 10.31234/osf.io/syp5a

Arslan, R.C., Schilling, K.M., Gerlach, T. M., & Penke, L. (in press). Using 26 thousand diary entries to show ovulatory changes in sexual desire and behavior. *Journal of Personality and Social Psychology*. Doi: 10.1037/pspp0000208

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social psychology*, *51*, 1173.

Blake, K. R., Dixson, B. J., O'Dean, S. M., & Denson, T. F. (2016). Standardized protocols for characterizing women's fertility: A data-driven approach. *Hormones and Behavior*, *81*, 74-83. Doi: /10.1016/j.yhbeh.2016.03.004

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365

Chambers, C.D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex, 49*, 609-610.

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. doi: 10.1016/j.psyneuen.2018.04.007

Fales, M. R., Gildersleeve, K. A., & Haselton, M. G. (2014). Exposure to perceived male rivals raises men's testosterone on fertile relative to nonfertile days of their partner's ovulatory cycle. *Hormones and Behavior*, *65*, 454-460. Doi: 10.1016/j.yhbeh.2014.04.002

Gangestad, S. W., Dinh, T., Grebe, N. M., Del Giudice, M., & Emery Thompson, M. (this issue). Psychological cycle shifts redux: Revisiting a preregistered study examining preferences for muscularity. *Evolution and Human Behavior*. Doi: https://doi.org/10.1016/j.evolhumbehav.2019.05.005

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,* 85-96. doi: 10.1016/j.evolhumbehav.2015.09.001

Gangestad, S. W., Garver-Apgar, C. E., Simpson, J. A., & Cousins, A. J. (2007). Changes in women's mate preferences across the ovulatory cycle. *Journal of Personality and Social Psychology*, *92*, 151.

Gangestad, S. W., Thornhill, R., & Garver-Apgar, C. E. (2005). Adaptations to ovulation: Implications for sexual and social behavior. *Current Directions in Psychological Science, 14*, 312–316. doi: 10.1111/j.0963-7214.2005.00388.x

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014). Do women's mate preferences change across the ovulatory cycle? A meta-analytic review. *Psychological Bulletin, 140*, 1205–1259. doi: 10.1037/a0035438

Higham, J. P. (2016). Field endocrinology of nonhuman primates: past, present, and future. *Hormones and Behavior, 84*, 145-155. Doi: 10.1016/j.yhbeh.2016.07.001

Huhtaniemi, I. T., Tajar, A., Lee, D. M., O'Neill, T. W., Finn, J. D., Bartfai, G., ... & Kula, K. (2012). Comparison of serum testosterone and estradiol measurements in 3174 European men using platform immunoassay and mass spectrometry; relevance for the diagnostics in aging men. *European Journal of Endocrinology*, *166*, 983-991.

Jones, B. C., Hahn, A. C., & DeBruine, L. M. (2019). Ovulation, Sex Hormones and Women's Mating Psychology. *Trends in Cognitive Sciences*, 23, 51-62. Doi: 10.1016/j.tics.2018.10.008

Jones, B. C., Hahn, A. C., Fisher, C. I., Wang, H., Kandrik, M., Han, C., … DeBruine, L. M. (2018). No compelling evidence that preferences for facial masculinity track changes in women's hormonal status. *Psychological Science, 29*, 996-1005. doi: 10.1177/0956797618760197

Jünger, J., Kordsmeyer, T. L., Gerlach, T. M., & Penke, L. (2018a). Fertile women evaluate male bodies as more attractive, regardless of masculinity. *Evolution and Human Behavior, 39*, 412-423. DOI: 10.1016/j.evolhumbehav.2018.03.007

Jünger, J., Motta-Mena, N. V., Cardenas, R., Bailey, D., Rosenfield, K. A., Schild, C., Penke, L., & Puts, D. A. (2018b). Do women's preferences for masculine voices shift across the ovulatory cycle? *Hormones and Behavior, 106*, 122-134. Doi: 10.1016/j.yhbeh.2018.10.008

Kordsmeyer, T., Hunt, J., Puts, D. A., Ostner, J., & Penke, L. (2018). The relative importance of intra- and intersexual selection on human male sexually dimorphic traits. *Evolution and Human Behavior, 39*, 424-436. DOI: 10.1016/j.evolhumbehav.2018.03.008

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*, 633–639. doi: 10.1016/j.yhbeh.2007.03.006

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97*, 951–966. http://doi.org/10.1037/a0028380

Marcinkowska, U. M., Galbarczyk, A., & Jasienska, G. (2018a). 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. doi: 10.106/j.psyneuen.2017.11.007

Marcinkowska, U. M., Kaminski, G., Little, A. C., & Jasienska, G. (2018b). Average ovarian hormone levels, rather than daily values and their fluctuations, are related to facial preferences among women. *Hormones and Behavior, 102,* 114-119. Doi: 10.1016/j.yhbeh.2018.05.013

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,* 38-43. doi: 10.1016/j.biopsycho.2005.02.002

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,* e4138. doi: 10.1371/journal.pone.0004138

Price, M. E., Dunn, J., Hopkins, S., & Kang, J. (2012). Anthropometric correlates of human anger. *Evolution and Human Behavior, 33,* 174-181. Doi: 10.1016/j.evolhumbehav.2011.08.004

Roney, J.R. (2018). Functional roles of gonadal hormones in human pair bonding and sexuality. In O. C. Schultheiss & P. H. Mehta (Eds.), *Routledge International Handbook of Social Neuroendocrinology* (pp. 239–255). New York, NY: Routledge

Roney, J. R., & Simmons, Z. L. (2016). Within-cycle fluctuations in progesterone negatively predict changes in both in-pair and extra-pair desire among partnered women. *Hormones and Behavior, 81,* 45-52. doi: 10.1016/j.yhbeh.2016.03.008

Schultheiss, O. C., Dlugash, G., & Mehta, P. H. (2019). Hormone measurement in social neuroendocrinology: A comparison of immunoassays and mass spectroscopy methods. In O. C. Schultheiss & P. H. Mehta (Eds.), *Routledge International Handbook of Social Neuroendocrinology* (pp. 26-40). New York, NY: Routledge

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9,* 666-681. Doi: 10.1177/1745691614553988

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11,* 702-712. Doi: https://doi.org/10.1177/1745691616658637

Stern, J., Gerlach, T. M., & Penke, L. (2019). Probing ovulatory cycle shifts in women's mate preferences for men's behaviors. *Preprint on PsyArXiv.* Retrieved from https://psyarxiv.com/7g3xc. Doi: https://dx.doi.org/10.17605/OSF.IO/7G3XC

van Stein, K. R., Strauß, B., & Brenk-Franz, K. (2019). Ovulatory Shifts in Sexual Desire But Not Mate Preferences: An LH-Test-Confirmed, Longitudinal Study. *Evolutionary Psychology, 17,* 1-10. Doi: https://doi.org/10.1177/1474704919848116

Wood, W., Kressel, L., Joshi, P. D., & Louie, B. (2014). Meta-analysis of menstrual cycle effects on women's mate preferences. *Emotion Review, 6,* 229-249. doi: 10.1177/1754073914523073