Evolution and Human Behavior 40 (2019) 501–516

Contents lists available at ScienceDirect

Evolution and Human Behavior

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

Psychological cycle shifts redux: Revisiting a preregistered study examining preferences for muscularity

Steven W. Gangestad[a,*], Tran Dinh[a], Nicholas M. Grebe[b], Marco Del Giudice[a], Melissa Emery Thompson[c]

[a] Department of Psychology, University of New Mexico, United States of America
[b] Department of Evolutionary Anthropology, Duke University, United States of America
[c] Department of Anthropology, University of New Mexico, United States of America

ABSTRACT

Jünger et al. (2018) conducted a preregistered study examining whether women particularly prefer muscular bodies when conceptive in their cycles. Despite an impressive number of participants and within-woman observations, they found no evidence for a preference shift; rather, they claimed, conceptive women find all male bodies more attractive. We preregistered a separate study very similar to Jünger et al.'s, with specified analyses focusing on shifts associated with joint additive effects of log-transformed estradiol and progesterone (ln(E/P)). We performed similar analyses on Jünger et al.'s publicly available data, using an empirically vetted (though not preregistered) measure of Strength/Muscularity. They revealed a ln(E/P) × Strength/Muscularity × Relationship Status interaction effect on sexual attraction. The ln(E/P) × Strength/Muscularity interaction ran in opposite directions for partnered and single women effects largely driven by P levels. Jünger et al.'s null conclusions and claims about general preferences are premature. We offer several observations regarding preregistered analyses.

1. Introduction

1.1. Cycle shifts

Do women's sexual interests change across the ovulatory cycle? If so, how? These questions have received tremendous attention over the past two decades. Findings converge on some answers. On average, during the peri-ovulatory phase, women become increasingly interested in sex and sensitive to stimuli evoking sexual motivation (e.g., Arslan, Schilling, Gerlach, & Penke, in press; Jones et al., 2018a; Roney & Simmons, 2013)—shifts likely mediated by changes in ovarian hormone levels (estradiol and progesterone; e.g., Roney & Simmons, 2013, found that, with ovarian hormone levels controlled, there was no significant residual effect of estimated conception risk). In other respects, answers remain elusive and theoretical issues unresolved. E.g., do partnered women become especially more attracted to men other than primary partners during the peri-ovulatory phase (e.g., Grebe, Emery Thompson, & Gangestad, 2016), or are increases in sexual attraction to both primary partners and other men similar (e.g., Roney & Simmons, 2016; Jones et al., 2018b; see also Dinh, Pinsof, Gangestad, & Haselton, 2017)?

A domain producing inconsistent results concerns mate preferences. Do women become increasingly attracted to some men, but not others, during the peri-ovulatory phase? Two meta-analyses of a sizable literature offer contrasting conclusions: one revealed an overall increase in attraction to a targeted set of male features during the peri-ovulatory phase (male facial masculinity, body masculinity, vocal masculinity, scent associated with developmental stability, features associated with greater male testosterone; Gildersleeve, Haselton, & Fales, 2014a); the other detected no such effects (Wood, Kressel, Joshi, & Louie, 2014; cf. Gildersleeve, Haselton, & Fales, 2014b).

Based on additional meta-analytic analyses, Gangestad, Grebe, Gildersleeve, and Haselton (2018) proposed that shifts in preferences may exist for some features (e.g., behavioral intrasexual competitiveness) but not others (e.g., facial masculinity, facial symmetry; see also Jones et al., 2018a, 2018b). Still, they emphasize, more research is needed. Among promising candidates for cycle shifts are preferences for muscular features. Jünger, Kordsmeyer, Gerlach, and Penke (2018); hereafter, Jünger et al.) empirically tested this possibility, as reported in *Evolution and Human Behavior*.

Jünger et al.'s study is truly impressive. Naturally ovulating women's preferences (N = 157) were assessed across four lab sessions and two cycles: twice during the peri-ovulatory phase, twice during the luteal phase. Peri-ovulatory status was assessed by luteinizing hormone (LH) tests (~90% positive). Women evaluated 80 digitally scanned male bodies represented in a rotating 3D format, stripped of distractions such as skin tone and heads. Steroid hormone levels, including estradiol and progesterone, were measured in saliva collected during every

[*] Corresponding author at: Department of Psychology, University of New Mexico, Albuquerque NM 87131, United States of America.
E-mail address: sgangest@unm.edu (S.W. Gangestad).

https://doi.org/10.1016/j.evolhumbehav.2019.05.005
Received 12 June 2018; Received in revised form 14 May 2019; Accepted 20 May 2019

session.

Jünger et al. examined changes in women's preferences for 6 male features argued to reflect muscularity/masculinity (see below), plus height; multilevel regression analyses failed to detect preference shifts across conceptive and non-conceptive phases for any of these features. The authors conclude, "Contrary to previously reported findings, men's masculine body characteristics did not interact with cycle phase to predict sexual attractiveness, indicating *no shifts in preferences for specific traits*" (p. 419; emphasis added). Instead, Jünger et al. emphasized a generalized cycle shift: in the peri-ovulatory phase, women rated *all* male bodies as more attractive on average—both as sex partners and long-term mates, and regardless of bodily features. Jünger et al. argue that this shift—highly robust in their analyses—is fully carried by partnered (vs. single) women.

1.2. Preregistration

One additional element of Jünger et al.'s study is important: They preregistered their study on a public open science site (Open Science Framework; osf.org). Hence, the hypotheses, study design, recruitment strategies, data-collection stopping rules, and data analytic strategies were planned out ahead of time and "announced." In light of psychology's replication crisis (e.g., Open Science Collaboration, 2015), for many scholars, this feature warranties the study's other admirable qualities. When unconstrained by a pre-announced plan, researchers have data analytic degrees of freedom (e.g., Simmons, Nelson, & Simonsohn, 2011). They may even modify, post hoc, the precise hypotheses tested to permit reporting of "positive" results (e.g., Gelman & Loken, 2013). While researchers may sincerely seek to understand their data through these practices (Simmons et al., 2011), the effects are insidious. False-positive rates and estimates of effects become inflated, hence littering the literature with non-replicable findings. Indeed, some scholars argue that these practices explain why some mate preference shifts have not replicated (e.g., Harris, Pashler, & Mickes, 2014).

Preregistration clearly serves a valuable function: By closing out researcher degrees of freedom, it controls α, the false-positive rate. By itself, however, preregistration does not guarantee meaningful results. Scholars must critically evaluate how results speak to theory, given how predictions were derived and analyses conducted. A non-controversial example makes the point: If a study design confounds a predictor variable with another variable, associations with the predictor remain ambiguously interpretable, regardless of whether the design is preregistered. In recognition of this point, some leading journals in psychology (e.g., *Psychological Science* [Lindsay, 2017]; *Journal of Personality and Social Psychology*) agree to report the results of a preregistered replication study, contingent on the preregistration passing stringent review prior to data collection. (See, e.g., https://cos.io/rr/.) A more basic question is whether preregistration should constrain authors to disregard additional evidence contradicting the findings of planned analyses.

1.3. The current paper

The current paper presents a critique and reanalysis of data from Jünger et al.'s published study. Some of us recently preregistered a study very similar to Jünger et al.'s, with detailed analyses that differ, in important ways, from Jünger et al.'s. While Jünger et al. focused on preference shifts according to cycle phase—which implies that hormonal mediators could be responsible—our analysis focuses directly on ovarian hormones as predictors of attraction to muscular features. We also address several confounds suggested by the outcomes of their data analysis. Thanks to Jünger et al.'s open data sharing, we were able to perform these analyses on their publicly available data. Empirical patterns contrast, in some ways sharply, with their claims. We explain how and why results importantly differ and can lead to different conclusions. Additionally, we illustrate broader points regarding preregistration with this study as example.

2. Jünger et al.'s analyses

In a general manner, Jünger et al.'s preregistration states hypotheses to be tested and suggests variables to be included in hypothesis tests. Specific statistical models, however, were absent from the preregistered document. Under "Statistical Models" of their online preregistration, Jünger and Penke (2016) write,

> Data will be analyzed using full-data multilevel modelling and lens models (Nestler & Back, 2013), … [S]exual and long-term attractiveness ratings serve as outcomes. The ovulatory cycle phase, measured steroid hormones, relationship status, LH ovulation test significance, personality traits, all cues specified in the hypotheses, latent variables as well as the relationship between hair hormone levels and average saliva hormone levels within and between women, will serve as predictors. [p. 7].[1]

A second paragraph lists confounding variables to be controlled. But substantial room for analytic flexibility remains (e.g., the preregistration itself does not specify how hormonal mediation will be evaluated). We describe the analytical decisions Jünger et al. presented.

2.1. Analysis of within-cycle shifts based on LH tests

In their preregistration, Jünger and Penke (2016) state, "Previous research has documented ovulatory cycle shifts in naturally cycling women that are assumed to be regulated by steroid hormonal changes (primarily by estradiol and progesterone)" (p. 3). As emphasized in their preregistration, key research questions addressed by their study were "Do naturally cycling women evaluate men differently for short-term relationships in their fertile window, relative to their non-fertile days? Do ovulatory cycle shifts on females' preferences of men's body masculinity, voice masculinity and socially flirtatious behavior exist?" and "Are menstrual cycle shifts in preferences mediated by changes in steroid hormones?" (Jünger & Penke, 2016, p. 3) They hence preregistered the hypotheses that "naturally cycling women in their fertile window, compared to their luteal phase, evaluate masculine stimuli (bodies, […]) as more attractive for short-term relationships", and that "the effect is mediated by a high estradiol and a low progesterone level" (p. 4). Hormone levels, if functioning as mediators, should predict changes in women's psychological states across the cycle better than estimated conception risk does—meaning analyses using hormonal predictors should have greater power. But despite having E and P levels available, Jünger et al. did not examine hormonal associations with preferences. Instead, they used estimated cycle phase as a predictor.[2]

[1] Hypotheses not tested by Jünger et al. correspond to mentions of lens models and hair hormones.

[2] Of course, physiological signals other than estradiol and progesterone *could*, in principle, be responsible for effects across conceptive and non-conceptive phases. Yet (a) no evidence points to particular candidates (see, e.g., Roney & Simmons, 2013, 2017, who found that, after estradiol and progesterone levels were controlled, cycle phase had no effect on sexual desire and food intake, respectively), and (b) Jünger and Penke (2016) did not preregister any other candidates, or suggest "partial" mediation by steroid hormones; the sole mediators they preregistered were steroid hormones. Indeed, the title of their preregistration was "The effects of ovulatory cycle shifts in *steroid hormones* on female mate preferences…" (emphasis added).

In a review of this commentary, Lars Penke, along with Julia Jünger and Ruben Arslan, claimed that this hypothesis concerning mediation by estradiol and progesterone only referred to main effects of cycle phase. They claimed that the hypothesis had nothing to do with *preferences* for masculine stimuli and, hence, the hormonal mediation hypothesis had nothing to do with preferences. We refer readers to supplementary online materials (SOM, section 26) for in-depth discussion of reasons why these claims about their preregistration are problematic.

2.2. Six male features putatively reflecting upper-body strength plus height

Jünger and Penke (2016) specifically preregistered the hypothesis that, when conceptive in their cycles, women will experience increased attraction to "*visual cues of upper-body strength* (e.g. shoulder-chest ratio, shoulder-hip ration [*sic*], upper-torso volume relative to lower-torso volume, upper arm circumference controlling for BMI)" (pp. 4–5; emphasis added). In addition to these 4 visual cues, Jünger and Penke (2016) preregistered hypotheses regarding preference shifts for physical strength, assessed in-lab, and male baseline testosterone level. They also preregistered the hypothesis that, when conceptive, women prefer taller male bodies. At the same time, Jünger et al. offered no evidence or justification for how features reflected upper body strength.

2.3. Simultaneous entry

In multilevel analyses, Jünger et al. regressed male sexual attractiveness on main effects for the 6 features and height, plus interactions between the features and cycle phase (see their Table 2). The 7 interaction terms constituted tests of cycle shifts: Cycle Phase × Strength, Cycle Phase × Arm Circumference, Cycle Phase × SHR, etc. None were statistically robust.[3]

It would be surprising if putative indicators of upper body strength did not covary. In Jünger et al.'s data, shoulder-to-chest ratio and shoulder-to-hip ratio covary strongly, probably because both variables share shoulder breadth as the numerator, $r = .64$. Strength and upper arm circumference also covary: $r = .50$. These indicators tap a common factor, unsurprisingly: muscular upper arms contribute to upper-body strength. If two interaction terms to assess preference shifts are entered—Cycle Phase × Strength and Cycle Phase × Arm Circumference—the analysis can only detect shifts in preference *uniquely* associated with each feature, *independent* of the other (i.e., strength *holding arm circumference constant*, arm circumference *holding strength constant*; Kutner, Nachtsheim, & Neter, 2004). Accordingly, the analysis is not especially sensitive to detecting shifts in preferences for the common factor. Suppose, for instance, a common factor generates a correlation of .5 between two equally-valid indicators, and an outcome covaries with the common factor. If power to detect an association of the outcome with a composite measure is 80% in a multiple regression, power to detect an association with an individual measure is just 29%.[4] In footnoted follow-up analyses, Jünger et al. regressed attraction on each male feature and its interaction with cycle phase individually, which they presented in supplementary online materials (SOM).

2.4. Control for main effects of a confounding feature (BMI)

Some "muscular" features highly covary with confounding non-muscular (indeed, unattractive) features. Most notably, r between bodies' upper arm circumference and body mass index (BMI) is .77. Men with well-developed musculature possess large upper arms, but so too do men with large fat depots. Arm circumference as a measure of muscularity, then, is contaminated by associations with fat. Strength too covaried with BMI, $r = .42$. Accordingly, Jünger et al. controlled for the *main* effect of BMI in analyses, which did not affect results.

However, Jünger et al. did not control for BMI confounding with *preference shifts*. Entering the main effect of BMI eliminates nuisance variance in attractiveness associated with BMI, by separating out BMI's confounding effects from a male feature's *main effect*. Yet it does nothing to control for BMI confounding with the primary effects of interest, those reflecting preference shifts. A Cycle Phase × Male Feature interaction is not confounded with the main effect of BMI; it is confounded with Cycle Phase × BMI. To fully control for these confounds, then, one must include a set of interaction terms with BMI paralleling interaction terms with a male feature. Alternatively, one can regress the male feature on BMI and compute residual scores, unconfounded with BMI, and use those in place of the male feature in analyses. As we quoted earlier, Jünger and Penke's (2016) explicitly preregistered a measure of "upper arm circumference controlling for BMI" (p. 4). That description implies a measure of residuals of upper arm circumference, with BMI controlled. Yet Jünger et al.'s analyses did not use this measure.

2.5. Consideration of relationship status

Jünger and Penke (2016) preregistered the hypothesis that "Cycle phase shifts in preferences for short-term mates are larger for partnered women than for single women" (p. 7; see also Hypothesis 4a, Jünger et al.; see, e.g., Havlicek, Roberts, & Flegr, 2005, cited by Jünger et al.). Statistically, analyses testing this hypothesis may examine whether Cycle Phase × Male Feature interactions are moderated—i.e., whether 3-way interactions exist: Cycle Phase × Strength × Relationship Status, Cycle Phase × Arm Circumference × Relationship Status, etc. But these analyses were not performed. Once Jünger et al. identified their primary positive finding from initial analyses—main effects of Cycle Phase on attraction—they dropped interaction terms involving male features. They only examined the role of relationship status, then, by assessing whether it moderates these main effects of cycle phase—e.g., whether Cycle Phase × Relationship Status effects are robust. Again, they argued yes. They did not examine whether relationship status moderates *cycle shifts in preferences for male features*—a key preregistered question of interest.

2.6. Summary

Jünger et al. made a number of analytic choices that can be reasonably debated. In particular, they chose four putative visual cues of upper-body strength without checking if they actually reflected strength, and—in their main analysis—entered them simultaneously as predictors (together with physical strength measured in the lab, testosterone, and height); this amounts to testing the unique effects of each feature, net of the common factor they were supposed to index (i.e., upper body strength). In addition, they deviated from their pre-registration in three ways. First, they only analyzed within-cycle preference shifts based on conceptive status (fertile vs. non-fertile) assessed with LH tests, despite having hypothesized that the effects would be mediated by estrogen and/or progesterone and having listed those variables in the pre-registration. Second, they did not control for the confounding effects of BMI on preference shifts for cues of upper body strength; this would have required including interaction terms in addition to the main effects of BMI. Third, they pre-registered the hypothesis of a 3-way interaction between cycle phase, upper body strength, and relationship status, but did not test this hypothesis in their analysis.

3. Alternative analyses

Gangestad et al. (2018) preregistered a now-ongoing study with similar study design features as in Jünger et al. (See https://osf.io/kd5j7/.) Women (N = ~250) arrive for 4 lab session assessments. They rate the sexual attractiveness of male bodies on multiple occasions. Peri-ovulatory sessions will be confirmed with LH tests. On the day of each session, women's biological samples will be collected for ovarian hormone assays. In several respects, however, our preregistered analysis plan differs from Jünger et al.'s, and in ways that pertain to our

[3] They regressed women's rated attraction for long-term relationships on male features too, but their primary preregistered hypothesis concerned sexual attraction.

[4] We assessed this in G*Power across true correlations of the common factor with an outcome ranging from 0.15 to 0.35; a near-identical drop in power occurred.

criticisms of their analyses.[5]

3.1. Primary analyses concern hormonal associations

Jünger et al. chose to focus primary analyses on session type (fertile vs. non-fertile), based on scheduling (using counting methods) and LH testing. By contrast, our primary analyses will examine associations with hormone levels. The reason is straightforward: If hormone levels drive variations across the cycle, as researchers commonly believe (e.g., Roney & Simmons, 2013) and Jünger and Penke (2016) preregistered, hormones should predict outcomes more strongly than conceptive status does. Even among healthy women of prime reproductive age, relative levels of ovarian hormones vary considerably across women and across cycles within the same woman, which moderate the likelihood that ovulation or conception will occur (Ellison, 2003; Lipson & Ellison, 1996). The regularity of menstrual cycles is not a guarantee of conceptive cycles. Even when precisely determined, the equivalent cycle day may have a dramatically different hormonal output (Ellison, 1993). And notably, women's days of participation within specific phases are not perfectly matched. Some are tested on a day of peak estradiol or progesterone, others days before or after it. Analyses using hormone levels are sensitive to these variations; analyses that categorize sessions as conceptive or non-conceptive are not. In our preregistration, analyses using LH-confirmed conception status as a predictor are secondary, not primary, analyses.[6]

In multilevel analyses, one can enter two orthogonal measures of variation for each hormone: within-woman (levels mean-centered within-woman); and between-woman (variation across woman-specific means; see West, Ryu, Kwak, & Chan, 2011). One might think that between-woman variation reflects individual differences or variation across cycles. While true if hormone levels are assayed daily (e.g., Roney & Simmons, 2013), when hormone levels are assayed sparingly across a cycle, much "mean" variation simply reflects when levels were assayed and not true differences across women or cycles. (I.e., even if every woman's cycle had identical hormone profiles, some "between-woman" variation would emerge, simply due to sampling at different points within the cycle.) Indeed, Cronbach's α of mean ln(E/P) in Jünger et al.'s data is just .22 (mean r across 4 measurements = .09), consistent with most variation in means reflecting within-woman, not between-woman, variation. Moreover, a reasonable assumption is that hormones have similar effects on outcomes, whether within-woman or between different women. Grand-mean centering hormone levels (as opposed to within-woman mean centering) allows for analysis of the total association of a hormonal measure with an outcome (e.g., Kreft, de Leeuw, & Aiken, 1995). We proposed to run both sets of analyses.

3.2. Log-transforming hormone levels and using the estradiol:progesterone ratio

In analyses examining outcome features in relation to hormonal predictors, log-transformation of hormone values is a common practice (Jones, 1996). Though transformation typically creates a distribution closer to normal, this is not the primary reason for transformation. Log-transformation changes the linearity of associations with other variables. Given how hormones affect outcomes—by binding to available receptors that diminish in availability as hormone levels rise—-hormonal effects often increase linearly with proportionate (i.e., log-transformed), not absolute, changes (Jones, 1996).

We specifically preregistered analyses examining outcomes (e.g., preference shifts) as a function of the log of the estradiol to progesterone ratio [ln(E/P)]. While E increases both prior to and after predicted ovulation, P is only produced in appreciable levels after ovulation. Furthermore, the two hormones have known antagonistic effects on sexual behavior (Dixson, 2013; Roney & Simmons, 2013). Thus, E/P is a biomarker of conceptive status (Baird, Weinberg, Wilcox, & McConnaughey, 1991), which, log-transformed, is ln(E/P). Ln(E/P) reflects simple additive effects of ln(E) and ln(P), as ln(E/P) = ln(E) – ln (P). Hence, in regression analyses, ln(E/P) captures equal but opposite joint additive contributions of ln(E) and ln(P). (It constrains the regression weights of ln(E) and ln(P) to be identical in magnitude but opposite in sign. E/P does not have a similar interpretation; see Sollberger & Ehlert, 2016.[7]) Joint but opposite effects can be detected with greater power using ln(E/P) than two separate predictors. Follow-up analyses entering ln(E) and ln(P) separately are necessary to evaluate unique contributions.[8]

At the same time, testosterone (T) levels may also affect outcomes (e.g., Welling et al., 2007) and covary with E and/or P. We control for these effects by also entering ln(T) and interactions paralleling ln(E/P) interactions. While female sexual behavior has also been attributed to T, its independent effects have been questioned (Wallen, 2013). Robustness analyses can assess the impact of removing ln(T) from the model. Grebe et al. (2016) applied analyses very similar to these to examine hormonal associations with in-pair and extra-pair sexual interests.

3.3. Muscular variation captured with a single measure

In our preregistered replication study, we use images of bodies that, as confirmed by pretesting, differ in musculature. A measure of third-party rated muscularity will be used as a predictor in analyses. By contrast, Jünger et al. presented an array of bodies exhibiting natural variation in muscularity; they used multiple bodily measurements, purportedly representing "upper body strength," as predictors in analyses. In their main analysis, Jünger et al. simultaneously entered the multiple putative indicators of upper body strength, compromising

[5] This preregistration was finalized and submitted to Open Science Framework on April 18, 2018. It was originally submitted for review to a journal (for purposes of a preregistered publication) in early February 2018. Jünger et al.'s data was made publicly available in January 2018, and we downloaded their data in mid-March 2018. Our preregistration (including fundamental priority of hormonal predictors, and treatment of all hormone levels, e.g., log-transforming the E/P ratio and using it as a primary predictor) follows a plan described in a grant proposal submitted to (January 2017) and ultimately funded (August 2017) by National Science Foundation.

[6] In fact, in 5% of the instances in which Jünger et al. could confirm an LH surge, women's "high fertility" session was conducted 3+ days after the surge. In another 9%, it was conducted 2 days after the surge, and in 12% it was conducted a day after the surge. Yet ovulation typically occurs less than a day following the LH peak (e.g., Wetzels & Hoogland, 1982); fertility has fallen dramatically (by 50–80%) even by the day of the LH peak (e.g., Dunson, Baird, et al., 1999, 2001). By day of ovulation, estradiol levels have dropped substantially (see Roney & Simmons, 2013, and references cited) and progesterone levels have begun to rise (e.g., Wetzels & Hoogland, 1982). In all likelihood, 10–20% of high fertility sessions in Jünger et al.'s sample (even among those with confirmed LH surges) were not conducted during a truly "high" fertility period, for timing reasons alone. (Additional ones could have been anovulatory. See Section 4.11.)

[7] Some researchers enter the untransformed E/P ratio into analyses, but interpretation is not straightforward. All variance in ln(E/P) is explained by simple additive effects of ln(E) and ln(P). By contrast, in Jünger et al.'s data, 20% of the variance in E/P is explained by additive effects of E and P, 4% by the linear E × P interaction, and 6% by E^2 and P^2. Over 70%, then, reflects complex non-linear main effects and interactions. In contrast to ln(E/P), E/P's meaning is unclear (see Sollberger & Ehlert, 2016, who broadly discourage use of raw hormone ratios; see also SOM, section 27).

[8] A reviewer wondered whether raw or logged hormone levels relate more strongly to conceptive status. In Jünger et al.'s sample with confirmed LH surges, both logged progesterone and the log of the E/P ratio predict "phase" (fertile vs. non-fertile) better than raw progesterone or the raw E/P ratio; $r = -0.60$, -0.73 for raw and logged progesterone values, respectively, and 0.38, 0.70 for raw and logged E/P ratios. The reviewer responded that this association may not generalize to other samples. See SOM, section 26, for further discussion of raw vs. log-transformed hormone measures and ratios.

Table 1
Jünger et al.'s data: sexual attractiveness and bodily dominance in relation to male features.

	Predicting sexual		Associations with		
	Attractiveness		Bodily dominance		
	γ/SE	t	p	*r*	*r* w BMI controlled
BMI	−.78/.2	−3.79	<**001**		
Strength	0.64/.20	3.17	**0.002**	.38***	.26*
BMI	−1.00/.29	−3.78	**0.001**		
Upper arm circumference	.65/.29	2.21	**0.03**	.51***	.35**
BMI	−0.59/.23	−2.54	**0.013**		
Shoulder-to-Chest ratio	−0.15/.23	−0.67	0.504	−.37***	−0.2
BMI	−.44/.21	−2.10	**0.039**		
Shoulder-to-Hip ratio	.16/.21	0.78	0.438	0.00	0.18
BMI	−.50/.20	−2.51	**0.014**		
Upper-to-Lower Torso Ratio	.06/.20	0.33	0.741	0.08	0.14
BMI	−.50/.20	−2.57	**0.012**		
Log Baseline Testosterone	.16/.19	0.82	0.417	0.07	0.08
	↑________________________↑ *r* between γ and partial *r* = .87				
BMI					
Height				−0.08	−0.2
BMI	−1.08/0.2	−4.35	<**.001**		
Factor: Strength/Arm Circ	.99/.25	3.43	**0.001**	.54***	.40**
BMI	−0.44/0.21	−2.08	**0.041**		
Factor: SCR/SHR	.18/.23	0.78	0.438	0.07	0.11
BMI	−0.48/0.2	−2.43	**0.017**		
Factor: Torso Ratio	.19/.24	0.73	0.466	0.08	0.17

Notes. Multilevel regression predicting sexual attractiveness from BMI and male feature. BMI and all features z-scored. Observations cross-classified by female raters (N = 157) and male targets (N = 80). Random intercepts for both modeled. Random slopes, across women, modeled for BMI and male features. Covariances between intercepts and slopes modeled. *df* for t = 77 to 83. N of male targets for correlations = 80. *** $p < .001$ ** $p < .01$ * $p < .05$. Confidence intervals are not explicitly reported. However, they can be very closely approximated with γ ± 2 × SE.
Note that, as γ for male feature increases, γ for BMI becomes more negative – likely because, when muscularity is controlled for, BMI becomes a "purer" measure of adiposity, which is unattractive. All p-values <.05 are in bold.

power to detect any one effect (though, as noted, they also included analyses entering individual features in their supplementary materials). Entering a single variable reflecting upper body strength, as reflected by multiple features aggregated into one measure, increases statistical power relative to entering multiple variables reflecting individual features (or single features one at a time). In our preregistration concerning preference shifts for behavioral displays, we capture behavioral variation with a single composite measure, an approach we recommend for analyzing Jünger et al.'s data.

Naturally, the indicator variable should validly reflect perceived upper body strength. Of the 6 male features potentially tapping upper body strength examined by Jünger et al., just one—strength—had a *main effect* on sexual attractiveness (see their Table 2). Yet prior research shows that women tend to find muscular bodies sexy, especially when unconfounded with fat (Frederick & Haselton, 2007; Millar, 2013). An obvious question arises: *Do these features truly reflect muscularity or upper body strength?*

We addressed this question in Jünger et al.'s dataset through a series of steps. First, we separately entered each male feature into a multilevel regression model predicting sexual attractiveness, controlling for BMI. Ratings were cross-classified by female participants, male targets, and their interaction, all for which we estimated random intercept variation. We also included random slopes for BMI and each male body feature to account for variation across women in impact of these features on ratings. Only Strength and Upper Arm Circumference significantly predict sexual attractiveness (all other *p*'s > .4). See Table 1.

Second, Kordsmeyer, Hunt, Puts, Ostner, and Penke (2018) asked men and women to rate these same 3-D scanned bodies on "Bodily Dominance"—how likely they were to win a physical fight. (Kordsmeyer et al. and Jünger et al. have overlapping authorship.) One can reasonably expect these ratings to reflect upper body strength, as well as overall size. With BMI controlled, Bodily Dominance was significantly and solely predicted by Strength and Upper Arm Circumference—the same features that predict sexual attractiveness; see Table 1. Consistent with muscularity being sexy, men's Bodily Dominance strongly predicts their mean sexual attractiveness to Jünger et al.'s women (BMI controlled), *r* = .73. The extent to which the 6 features correlate with Bodily Dominance strongly covaries with the extent to which they predict sexual attractiveness (BMI controlled), *r* = .87. See Table 1.

Third, we factor analyzed the 6 male features (principal axis extraction, direct oblimin rotation). A scree slope suggested 3 factors (eigenvalues = 2.23, 1.47, 1.01, .59, .43, .27). Strength and Upper Arm Circumference primarily define one factor (pattern matrix loadings of .71 and .73). Shoulder-to-Chest Ratio (−.38) and testosterone level (.34) have secondary loadings on this factor. Shoulder-to-Hip Ratio and Shoulder-to-Chest Ratio define a second factor (loadings of .84 and .67), and Torso Ratio (.80) a third. (See Table S1 in SOM for full loadings matrix.) Only the first factor relates to attractiveness or Bodily Dominance. See Table 1.

In sum, the empirical evidence converges on a clear conclusion: Two of the 6 features reflect muscularity; the others do not (at least not substantially).[9] Accordingly, we used a simple unit-weighted composite of Strength and Arm Circumference in our analyses. We refer to this composite score as Strength/Muscularity, though recognizing that this

[9] One can ask why the other 4 features don't reflect muscularity. Muscular men may have broad shoulders *and* chests, such that the ratio minimally covaries with muscularity. Shoulder-to-Hip and Torso Ratio might reflect small hips as much as than large upper bodies. Men's testosterone levels don't strongly predict muscular development (e.g., Alvarado et al., 2016). In any event, the evidence is clear: These features don't strongly reflect muscularity in Jünger et al.'s bodies.

composite does not fully capture muscularity and is conflated with fat mass (such that BMI must be controlled in statistical analyses, as we detail below). In our analyses, effects of primary interest contain a ln(E/P) × Strength/Muscularity component.[10]

Male height. Pawlowski and Jasienska (2005) found that, during the follicular phase compared to the luteal phase, women particularly preferred taller men. (A weakness of this study is that it did not examine the impact of fertility status per se.) Some scholars have argued that male height is associated with formidability (e.g., Fessler, Holbrook, & Snyder, 2012; Lukaszewski, Simmons, Anderson, & Roney, 2016), though evidence is mixed (see Sell et al., 2009). We subjected height to the same tests we submitted putative indicators of upper body strength. Independent of BMI, height did not predict attractiveness or Body Dominance (see Table 1). (The latter correlation was actually negative, though not significant, $r = -.20$, $p = .073$. The correlation without BMI controlled was near-zero, $r = -.08$.) In Jünger et al.'s sample, then, taller men were neither more attractive nor perceived to be more formidable. Male bodies shown to raters were headless, such that women could not perceive full height. Head size does not scale 1:1 with body size and, hence, smaller relative head size is a cue to height; raters lacked that cue of height as well. In any event, because height was not perceived as attractive or indicative of strength, we did not include it in analyses (except, as we note immediately below, as a component of BMI, which we controlled for).[11]

3.4. Control for preference shifts for confounding features

Men's BMI is highly confounded with their Strength/Muscularity ($r = .69$), meaning shifts in aversion to certain components of high BMI—e.g., “flabbiness”—are confounded with shifts in preference for Strength/Muscularity. To fully control for confounds with preferences, one must include a set of terms with BMI paralleling terms with Strength/Muscularity (e.g., ln(E/P) × BMI). Alternatively, one can regress Strength/Muscularity on BMI and compute residual scores, unconfounded with BMI, and use those in analyses. We analyzed results using both methods as a robustness check.[12]

Moderation by relationship status. To test moderation by relationship status, we include the ln(E/P) × Strength/Muscularity × Relationship Status interaction. This hypothesis had been specified in Jünger et al.'s pre-registration but was not tested in their analysis.

3.5. Summary

Our analyses contrast with Jünger et al.'s in a number of ways. We summarize major differences in Table 2.

4. Results

Below, we present our analyses and results of Jünger et al.'s data, downloaded from the Open Science Framework. We begin by presenting a model that fully reflects the analytic strategy we outline above and in our preregistration (Section 4.1). Next, we perform a series of robustness analyses based on this full model that examine how the exclusion of certain variables (Section 4.2), differing transformations of variables (Sections 4.3–4.4), and alternative operationalizations of predictor variables (Sections 4.8–4.10) affect results. In addition, we perform analyses that separately examine effects of estradiol and progesterone (Section 4.5), as well as estimate effects within partnered and single women separately (Sections 4.6–4.7). Table 3 describes the flow of these analyses. Both Jünger et al.'s and our preregistration emphasized moderation of impacts of bodily features on sexual attraction (vs. attraction to long-term mates). Hence, we focus on sexual attractiveness as a criterion. For completeness, we report analyses on attraction to men as long-term mates in Table S20.

4.1. Initial analysis

In our multilevel regression model, women's ratings of sexual attractiveness were cross-classified by female participants, male targets, and their interaction; random intercept variation was estimated for all. Predictors were within-woman ln(E/P), within-woman ln(T), woman-mean ln(E/P), woman-mean ln(T), Strength/Muscularity, BMI, and relationship status. Within-woman hormonal measures were zero-centered within-woman. Relationship status was effect-coded (single = −.5, paired = .5). All other measures were grand-mean zero-centered. Interactions involving a hormone level × male feature × relationship status (and all embedded 2-way interactions) were entered. Random slope variation across women was estimated for within-woman hormone levels, Strength/Muscularity, and BMI.[13] See our supplemental R markdown file (end of SOM) for R code used to run this and all other models.

Table 4 (full model) presents results. Most terms are control variables. Two are of primary interest: within-woman ln(E/P) × Strength/Muscularity and within-woman ln(E/P) × Strength/Muscularity × Relationship Status. The former did not emerge; the latter did ($p = .014$); hence, the two-way interaction was found to vary as a function of relationship status. As ln(E/P) increased, so too did partnered women's preference for Strength/Muscularity (see below), supporting Jünger et al.'s preregistered Hypothesis 4a.

A significant negative mean ln(E/P) × BMI × Relationship Status interaction also emerged. As partnered women's mean ln(E/P) increased, so too did their preference for lower BMI, independent of Strength/Muscularity. BMI independent of Strength/Muscularity likely reflects adiposity, in part, which might explain BMI's very robust negative main effect on attractiveness.[14]

For our own study, we will examine effects controlling for session number. Jünger et al. controlled for male age too, which may be confounded with muscularity. In Tables S4 and S7, we present analyses

[10] This composite correlates 0.97 with corresponding factor scores. In robustness analyses, we used factor scores, which yielded near-identical results. See Table S8.

[11] We factor analyzed height along with the 6 male features putatively indicative of upper body strength. Once again, one factor was defined most strongly by strength and upper arm circumference. Two other features had loadings that exceeded 0.5: height and shoulder-to-chest ratio (negatively, such that men with large chests relative to shoulder breadth had high factor scores). The factor, then, reflected size and strength, though, because height was not a cue of formidability in this sample of headless bodies, the correlation of factor scores for this factor with Bodily Dominance, independent of BMI, was relatively weak, $r = 0.20$, $p = .073$. As part of our robustness analyses, we substituted these factor scores (Strength/Muscularity/Height) for Strength/Muscularity. Analyses produced very similar findings and do not alter conclusions. Results are provided in Table S9; see also Figure S1, section 21.

[12] Including BMI effects in the analysis removes not only confounds but also nuisance variance in attraction associated with confounds. As well, it permits examination of BMI effects. For these reasons, we prefer it, though analysis using residual scores simplifies the model. Once again, Jünger et al.'s preregistration stated that upper arm circumference would control for BMI.

[13] Estimates may be sensitive to model selection: random intercept and slope terms. We used model fit statistics to select models. See S2 in SOM. Seven outlying hormone values, identified by visual inspection (2 progesterone, 5 testosterone; all values 2+ s from nearest retained value), were excluded. Their exclusion did not affect results. See Table S3 for analyses including these values.

[14] Reviewers questioned this interpretation, as relatively few bodies in Jünger et al.'s sample qualified as “overweight,” let alone obese. (10% of BMIs were > 26.) The variation in BMI in this sample, then, may not be meaningful. Extremes leverage correlations, however; 10% overweight individuals may well be enough to generate meaningful variation. And, indeed, BMI's very robust negative main effects (net of Strength/Muscularity) on attraction—effects as large of those of Strength/Muscularity—demand explanation; they betray the view that variation in BMI in this sample is not meaningful. In part, independent of muscularity, BMI must reflect adiposity.

Table 2
Key differences between our analyses and those of Jünger et al.

	Jünger et al.'s analyses	Our analyses
Purported drivers of shift entered in analyses	Estimated Cycle Phase	Measured hormone levels (notably, ln(E/P), as well as ln(E) and ln(P))
Male muscular features	6 features plus height entered simultaneously	A single composite, with components empirically vetted
Control for BMI confound	Controlled for main effect	Controlled for confounding BMI interactions
Test of moderation of preference shifts by relationship status	Did not test these interactions	Explicitly tested the ln(E/P) × Strength/Muscularity × Relationship Status interaction

Notes. The differences listed are primary ones. We note several additional differences: (a) Jünger et al. performed follow-up analyses (though not examining preference shifts) using raw hormone levels, not log-transformed levels; we performed robustness analyses with raw hormone levels that yielded the key ln(E/P) × Strength/Muscularity × Relationship Status interaction (see Table S10). (b) We eliminated some outlying hormone values through visual inspection; we performed robustness analyses with the full dataset that yielded the same key results (see Table S3). (c) We did not control for male age in the primary analyses; we performed robustness analyses including age that yielded the same key results (see Table S7). (d) We controlled for women's testosterone level (log-transformed) in primary analyses, whereas Jünger et al. did not; we also performed robustness analyses without controlling for ln(T) that yielded the same key results. (e) We included random slopes in our mixed model analyses, whereas Jünger et al. did not.

Table 3
Our analyses: An initial full model plus additional analyses examining robustness.

A full model (Table 4). We begin with a full model that follows from our overarching rationale. It uses ln(E/P) as a primary hormonal variable of interest, which has two orthogonal components, woman-mean and within-woman. The model also includes ln(T) as a control variable, which also has two orthogonal components. Strength/Muscularity is used as a marker of male muscularity. BMI is entered as a control variable. Relationship status is entered as a potential moderator. The primary effects of interest are within-woman ln(E/P) × Strength/Muscularity and within-woman ln(E/P) × Strength/Muscularity × Relationship Status. To control for preference effects of T and the confounding of preferences for BMI and Strength/Muscularity, however, 2-way interaction and 3-way interaction terms involving these variables must also be entered.

A model removing ln(T) (Table 4). We ran the same model as above, but removing ln(T) and all interactions. This analysis examines whether a simplified model not controlling for T yields the same effects.

Grand-centered mean analysis (Table 4). An analysis that grand-mean centers hormone values captures the total hormonal effects, both within and across women.

Strength/Muscularity residual scores, with BMI partialled out (Table 4). An alternative to entering BMI and its interactions is to regress Strength/Muscularity on BMI and use residual scores as a measure of Strength/Muscularity independent of BMI. We report this analysis using the grand-mean centered analysis approach described above.

Follow-up analyses examining separate contributions of ln(E) and ln(P) (Table 5). In these analyses, ln(T) is dropped, as (a) its inclusion introduces additional terms, and (b) robustness analyses described above show that its exclusion does not meaningful change key results.

Estimation of effects specific to partnered and single women (Table 6). In light of a ln(E/P) × Strength/Muscularity × Relationship Status effect, we follow up with analyses that separately examine the ln(E/P) × Strength/Muscularity effect within partnered and single women separately, using the grand-mean centered analysis described above. As well, we provide, for partnered women, model-based estimates of associations of ln(E/P) with sexual attraction to highly muscular and unmuscular men (95th and 5th percentile on Strength/Muscularity, respectively).

The SOM presents additional robustness analyses. The main text presents additional analyses using Bodily Dominance and a composite measure of Strength/Formidability as separate measures of muscularity (Table 7) and cycle phase as a potential driver of preference shifts (Table 9).

controlling for these features. Test-statistics for the within-woman ln(E/P) × Strength/Muscularity × Relationship Status effect are nearly identical (slightly stronger in each analysis).

4.2. Excluding ln(T) and between-woman terms

With ln(T) and its interactions (largely non-significant) excluded, the ln(E/P) × Strength/Muscularity × Relationship Status effect remains significant ($p = .019$). See Table 4. Within-woman and between-woman (woman-mean) hormonal terms are orthogonal and, hence, inclusion of the latter should not substantially affect estimation of the former. We did run analyses that excluded between-woman terms, both with and without ln(T) and its interactions included. As expected, the ln(E/P) × Strength/Muscularity × Relationship Status effects were nearly identical. See Table S5, SOM.

4.3. Estimating overall effects of ln(E/P)

Much "between-woman" variation in sampled E and P levels is, in fact, within-woman variation, arising from variable timing of sampling across women's cycles. But even if mean levels truly reflect between-woman variation (e.g., some women experience repeated anovulatory cycles), a parsimonious prediction is that equivalent concentrations of hormones produce similar responses, whether occurring in the same woman or different women. In such circumstances, entry of a grand-mean centered predictor (here, ln(E/P)) is the most powerful approach (e.g., Kreft et al., 1995). In this analysis, a positive ln(E/P) × Strength/Muscularity × Relationship Status interaction ($p = .005$) is significant. Among partnered women, high levels of ln(E/P) associate with increased preference for Strength/Muscularity. See Table 4.[15]

4.4. Using residual strength/masculinity scores

As expected, Strength/Muscularity residual scores (with BMI partialled out) yield very similar results. Table 4 presents a model (ln(T) terms excluded) retaining three predictors—ln(E/P), residual Strength/Masculinity, Relationship Status—and their interactions (hence, a fairly simple model with just 7 terms); 3-way interaction $p = .008$.

4.5. Estimating independent effects of ln(E) and ln(P)

The regression analyses above constrain ln(E) and ln(P) to have weights equal in magnitude but opposite in sign. In follow-up analyses we examined their independent effects. The effects of ln(P) are robust: ln(P) interacts (negatively) with Strength/Muscularity and Relationship Status to predict attraction; ln(E) does not. See Table 5 and Table S6.

4.6. Estimation of effects within partnered and single women

Assigning a value of zero to single or partnered women in relationship status coding, respectively, yields model-based estimates of all lower-order main effects and interactions for each group. The grand-mean centered ln(E/P) × Strength/Muscularity interaction is positive for partnered women, though it falls just short of statistical significance, $p = .061$. For single women, it significantly runs in a negative direction. See Table 6. See Table S17 for estimates separately examining within-

[15] For these analyses, 76% of total variation in ln(E/P) is explicitly within-woman. Again, a portion of between-woman variation is actually within-woman and arises as between-woman due to variable timing of sessions. All in all, the vast majority of total variance is within-woman.

Table 4
Results of multilevel regression analyses on jünger et al.'s data: Predictors of sexual attractiveness.

	Full Model[c]			T removed			GM centered E/P[b]			With residual S/M		
	γ/SE	*t*	*p*	γ/SE	*t*	*p*	γ/SE	*t*	*p*	γ/SE	*t*	*p*
BMI	−1.11/0.25	−4.48	<**.001**	−1.11/0.25	−4.46	<**.001**	−1.10/0.25	−4.38	<**.001**			
Strength/Muscularity (S/M)	.86/.25	3.51	<**.001**	.87/.25	3.50	<**.001**	.86/.25	3.47	<**.001**	.63/.19	3.34	**0.001**
Relationship Status	−0.01/0.1	−0.08		.25/.08	3.14	**0.002**	.11/.10	1.10		.31/.07	4.35	<**.001**
ww[a] E/P	.06/.04	1.58	0.117	.06/.04	1.91	*0.059*	.07/.04	1.74	*0.084*	.07/.04	1.87	*0.064*
ww T	−0.04/0.04	−0.87					−.06/.07	−0.77				
Mean E/P	−.04/.05	−0.78		.08/.08	0.96							
Mean T	.07/.08	0.79										
ww E/P x Relationship Status	.02/.06	0.39		.01/.06	0.13		−.02/.07	−0.37		−.03/.06	−0.57	
ww T × Relationship status	−0.22/0.08	−2.71	**0.008**				−.37/.10	−3.57	<**.001**			
Mean E/P × Relationship stat	.09/.10	0.82		−.11/.07	−1.67	*0.094*						
Mean T × Relationship status	−0.11/0.11	−0.93										
BMI × Relationship status	−0.09/0.06	−1.64	0.101	−.08/.06	−1.43	0.153	−.03/.05	−0.55				
BMI × ww E/P	−0.01/0.01	−0.54		−.00/.01	−0.44		−.01/.01	−0.65				
BMI × ww T	.02/.01	1.58	0.114				.03/.02	2.10	**0.036**			
BMI × mean E/P	−0.02/04	−0.58		−.02/.04	−0.47							
BMI × mean T	.06/.04	1.50	0.135									
S/M × Relationship Status	.03/.05	0.70		.03/.05	0.57		.03/.04	0.61		.02/.03	0.55	
S/M × **ww E/P**	−.00/.01	−0.29		−.00/.01	−0.34		−.00/.01	−0.15		−.00/.01	−0.34	
S/M × ww T	−0.01/0.01	−1.09					−.02/.02	−1.52	0.129			
S/M × mean E/P	.01/.03	0.28		.00/.02	0.13							
S/M × mean T	−0.03/0.03	−1.13										
Rel stat × BMI × ww E/P	−.02/.02	−1.22	0.222	−.02/.02	−1.09		−.04/.02	−1.78	*0.074*			
Rel stat × BMI × ww T	.03/.02	1.45	0.146				.02/.03	0.56				
Rel stat × BMI × mean E/P	−.16/.05	−3.26	**0.001**	−.16/.05	−3.24	**0.001**						
Rel stat × BMI × mean T	−0.05/0.05	−1.04										
Rel stat × **S/M** × **ww E/P**	.05/.02	2.47	**0.014**	.05/.02	2.34	**0.019**	.06/.02	2.78	**0.005**	.04/.02	2.65	**0.008**
Rel stat × S/M × ww T	−0.02/0.02	−1.16	0.246				−.00/.03	−0.12				
Rel stat × S/M x mean E/P	.06/.04	1.34	0.179	.06/.04	1.42	0.155						
Rel stat × S/M x mean T	.05/.04	1.09										

Notes. All hormone measures log-transformed. Hence, ln(E/P) = ln(E) - ln(P). All quantitative predictors z-scored. Relationship status effect coded: single = −.5, partnered = .5. Observations cross-classified by female raters ($N = 157$), male targets ($N = 80$), and their interaction. Random intercepts for all are modeled. Random slopes, across women, modeled for BMI, Strength/Muscularity, and within-woman hormone measures. Inclusion of random slope interactions and covariances selected through model Bayesian Information Criterion fit statistic. Random components and fit statistics reported in Table S2, SOM. Effects of primary theoretical interest **bolded**. Blank rows separate main effects, two-way interactions, and three-way interactions. *P*-values < .05 bolded. *P*-values < .10 in italics. *P*-values > .25 not shown. Confidence intervals are not explicitly reported. However, they can be calculated with γ ± 2 × SE.

[a] ww = within-woman centered.

[b] Grand-mean centered hormone measures reported in this table in rows for within-woman hormone measures.

[c] Strength/Muscularity scores regressed on BMI to remove confounding with BMI. Grand-mean centered hormone measures reported in rows for within-woman hormone measures.

Table 5
Results of multilevel regression analyses: Predictors of sexual attractiveness separating estradiol and progesterone.

	Full model			With residual S/M[a]		
	γ/SE	*t*	*p*	γ/SE	*t*	*p*
BMI	−1.11/0.25	−4.42	<**.001**			
Strength/Muscularity (S/M)	.86/.25	3.49	<**.001**	.64/.19	3.31	**0.001**
Relationship status	.02/.10	1.67	*0.096*	.16/.10	1.67	*0.095*
E	−.10/.08	−1.34	0.181	−.10/.08	−1.35	0.181
P	−.07/.03	−2.22	**0.029**	−.07/.03	−2.22	**0.029**
E × Relationship status	−.13/.12	−1.08		−.03/.12	−1.08	
P × Relationship status	.04/.05	0.68		.04/.05	0.69	
BMI × Relationship status	−.03/.05	−0.52				
BMI × E	−.02/.01	1.41	0.159			
BMI × P	.04/.06	0.91				
S/M × Relationship status	.03/.04	0.63		.00/.05	0	
S/M × **E**	−.02/.01	−1.58	0.114	−.01/.01	−1.46	0.145
S/M × **P**	−.00/.01	−0.25		−.00/.01	−0.19	
Rel stat × BMI × E	−.03/.03	1.2	0.229			
Rel stat × BMI × P	.05/.02	2.29	**0.022**			
Rel stat × **S/M** × **E**	.01/.03	0.38		.00/.02	0.23	
Rel stat × **S/M P**	−.06/.02	−2.75	**0.006**	−.04/.02	−2.74	**0.006**

Notes. Hormone values log-transformed and grand-mean centered. See also notes, Table 4. See S6 for full model analyses.

[a] Strength/Muscularity scores regressed on BMI to remove confounding with BMI.

Table 6
Results of multilevel regression analyses: Predictions for single and partnered women.

	Single			Partnered								
	Mean-Centered S/M			Mean-Centered S/M			S/M at 5th percent			S/M at 95th percent		
	γ/SE	*t*	*p*	γ/SE	*t*	*p*	γ/SE	*t*	*p*	γ/SE	*t*	*p*
Analysis with ln(E/P)												
BMI	−1.09/.25	−4.32	**<.001**	−1.11/.25	−4.44	**<.001**						
Strength/Muscularity (S/M)	.85/.25	3.42	**0.001**	.87/.25	3.52	**<.001**						
E/P	**.08/.05**	1.63	0.106	**.06/.05**	1.12		.02/.06	0.27		.11/.06	1.82	*0.070*
T	.13/.09	1.49	0.139	−.24/.09	−2.72	**0.007**						
BMI × E/P	.01/.02	0.79		−.03/.02	−1.74	0.083						
BMI × T	.02/.02	1.1		.04/.02	1.97	**0.049**						
S/M × E/P	−.03/.02	−2.05	**0.041**	.03/.02	1.87	*0.061*						
S/M × T	−.02/.02	−0.98		−.03/.02	−1.21	0.226						
Analysis with ln(E) and ln(P)												
E	−.04/.09	−0.42		−.17/.10	−1.68	*0.095*	−.14/.10	−1.30	0.195	−.20/.11	−1.90	*0.060*
P	−.09/.04	−2.19	**0.030**	−.05/.05	−1.21	0.229	−.00/.05	−0.08		−.11/.05	−2.14	**0.033**
BMI × E	.00/.02	0.16		.03/.02	1.78	*0.075*						
BMI × P	−.02/.02	−0.95		.04/.02	2.31	**0.021**						
S/M × E	−.02/.02	−1.45	0.148	−.02/.02	−0.83							
S/M × P	.03/.02	1.73	*0.084*	−.03/.02	−2.17	**0.030**						

Notes. Hormone values log-transformed and grand-mean centered. All quantitative predictors with $s = 1$. For Single estimates, relationship status coded Single = 0, Partnered = 1; for Partnered estimates, Single = 1, Partnered = 0. Interactions involving relationship status are redundant with Table 3 and Table 4 and are not shown. For analysis with ln(E) and ln(P), BMI and S/M main effects are not repeated. S/M at 5th percent = zero-centered at 5th percentile. S/M at 95th percent = zero-centered at 95th percentile. See S2 in SOM for discussion of random components. Effects of primary theoretical interest **bolded**. *P*-values < .05 bolded. *P*-values < .10 in italics. *P*-values > .25 not shown. Confidence intervals are not explicitly reported. However, they can be calculated with γ ± 2 × SE.

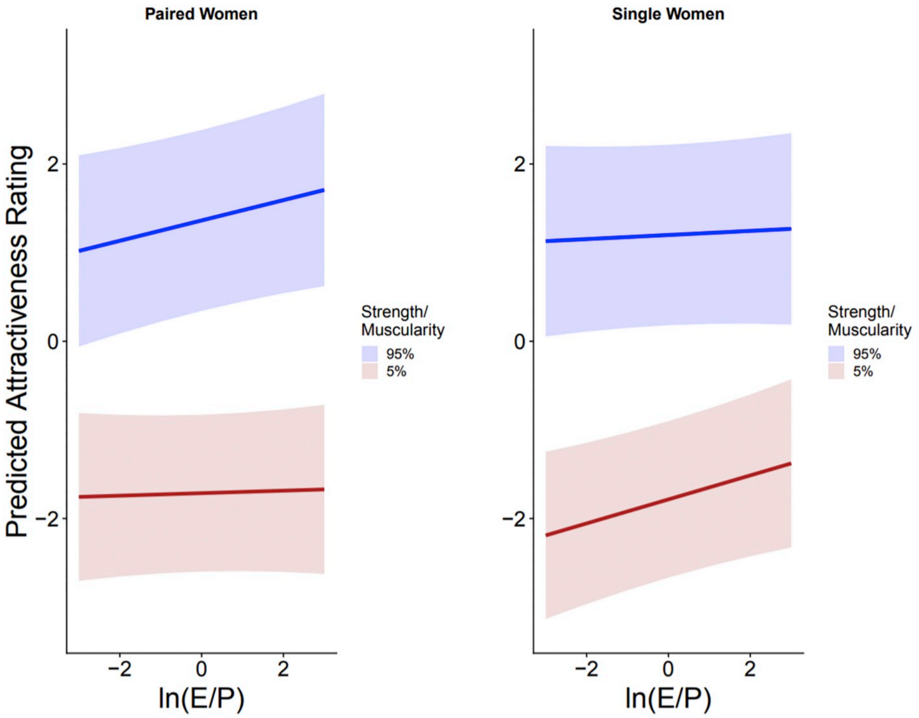

Fig. 1. Model-based estimates of the association between the log of E/P when Strength/Masculinity is at the 5th percentile and 95th percentile for partnered women (top panel) and single women (bottom panel). Shaded areas represent 95% confidence intervals.

woman and woman-mean hormone levels.

4.7. Estimation of preferences for high vs. low strength/muscularity men

With partnered women assigned a value of zero in relationship status coding and Strength/Muscularity zero-centered at the 5th and 95th percentiles ($z = -1.60$, 1.91, respectively), one derives model-based estimates of the effect of ln(E/P) on partnered women's attraction to highly unmuscular and very muscular men, respectively. See Table 6. As can be seen, partnered women's ln(E/P) positively predicts attraction to muscular men (though the effect falls just short of statistical significance, $p = .07$. It does *not* predict their attraction to non-muscular men, with effect size near-zero. Though no firm conclusions can be drawn, these results lead one to question Jünger et al.'s claim that,

Table 7
Results of multilevel regression analyses: predictors of attractiveness with bodily dominance and strength/formidability.

	Bodily dominance			Strength/Formidability		
	γ/SE	*t*	*p*	γ/SE	*t*	*p*
Analysis using E/P						
BMI	−1.06/0.15	−7.18	**<.001**	−1.47/.21	−7.06	**<.001**
BD/SF	1.39/.15	9.24	**<.001**	1.43/.21	6.94	**<.001**
Relationship Status	.10/.10	1.04		.11/.10	1.10	
E/P	.07/.04	1.77	*0.079*	.07/.04	1.74	*0.084*
T	−.06/.07	−0.77		−.06/.04	−0.77	
Relationship status × E/P	−.03/.07	−0.42		−.02/.07	−0.37	
Relationship status × T	−.38/.10	−3.59	**<.001**	−.37/.10	−3.58	**<.001**
BMI × Relationship status	- .04/.05	−0.92		−.06/.06	−1.02	
BMI × E/P	−.00/.01	−0.02		.01/.01	0.09	
BMI × T	.02/.01	1.40	0.162	.03/.02	1.78	*0.075*
BD/SF × Relationship status	.09/.05	1.73		.06/.05	1.14	
BD/SF × E/P	−.02/.01	−2.36	**0.018**	−.01/.01	−1.29	0.196
BD/SF × T	−.00/.01	−0.07		−.02/.02	−1.01	
Rel stat × BMI × E/P	−.02/.02	−1.17	0.24	−.05/.02	−0.19	
Rel stat × BMI × T	.01/.03	0.55		.02/.03	0.63	
Rel stat × BD/SF × E/P	.06/.02	3.25	**0.001**	.08/.02	3.54	**<.001**
Rel Stat x BD/SF x T	.00/.03	0.15		−.01/.03	−0.21	
Analysis entering E and P separately[a]						
E	−.10/.08	−1.32	0.188	−.10/.08	1.34	0.181
P	−.07/.03	−2.24	**0.027**	−.07/.03	−2.22	**0.029**
Relationship status × E	−.13/.11	−1.08		−.13/.12	−1.08	
Relationship status × P	.04/.06	0.75		.04/.06	0.68	
BMI × E	.02/.01	1.45	0.147	.02/.01	1.81	*0.071*
BMI × P	.00/.01	0.32		.00/.01	0.31	
BD/SF × E	−.03/.01	−2.68	**0.007**	−.03/.01	−2.29	**0.022**
BD/SF × P	.01/.01	1.52	0.130	.01/.01	0.63	
Rel stat × BMI × E	.03/.02	1.54	0.123	.03/.03	1.09	
Rel stat × BMI × P	.03/.02	1.78	*0.074*	.06/.02	2.72	**0.007**
Rel stat × BD/SF × E	.01/.02	0.5		.01/.03	0.52	
Rel stat × BD/SF × P	−.06/.02	−3.16	**0.002**	−.07/.02	−3.47	**<.001**

Notes. All hormone measures log-transformed and grand-mean centered. See notes, Table 3. BD = Bodily Dominance. SF = Strength/Formidability. Effects of primary interest **bolded**. *P*-values < .05 bolded. *P*-values < .10 in italics. *P*-values > .25 not shown. Confidence intervals are not explicitly reported. However, they can be calculated with γ ± 2 × SE. See Tables S14-S19 for full model analyses and effects for single and partnered women separately.

[a] For analyses entering E and P separately, for sake of brevity we do not repeat effects for main effects and interactions without E or P, though these terms were included; see the analysis using E/P.

when conceptive (or, here, when experiencing hormonal patterns reflective of fecundability), partnered women rate bodies *in general* as more sexually attractive, independent of men's bodily features. Effects for ln(P) are similar to those for ln(E/P) (but reversed in sign and, in the case of men at the 95th percentile, statistically significant, $p = .033$). These contrasting patterns are illustrated in Fig. 1.

4.8. Moderation of the association between bodily dominance and sexual attractiveness ratings

We used Kordsmeyer et al.'s (2018) ratings of Bodily Dominance to vet male features. Substituting Bodily Dominance for Strength/Muscularity is expected to produce similar results, as it likely reflects overall perceived muscularity, plus body size. And it does: a significant 3-way ln(E/P) × Bodily Dominance × Relationship Status interaction emerged ($p = .001$). See Table 7 and Table S14 and Fig. S2 (section 21). This 3-way interaction involving a separate (and raw, unprocessed) measure of male muscularity should bolster confidence in these effects' robustness. Bodily dominance ratings are completely distinct from any of the 7 male features and, hence, these effects do not depend on any particular composite of those features.

4.9. Moderation of strength/formidability and sexual attractiveness ratings

Strength, upper arm circumference, and Bodily Dominance covary considerably, $r = .38$–.51, all $p < .001$. A first principal component of all 3 (loadings of .78, .85, and .78, respectively) could be an even better measure of perceived muscularity. Component scores, which we call Strength/Formidability, covary almost perfectly with a unit-weighted sum ($\alpha = .72$; $r > .999$). Not surprisingly, in multilevel analyses, ln(E/P) interacts with Relationship Status and Strength/Formidability to predict sexual attraction, $p < .001$. See Tables 7 and S15 and Fig. S3 (section 21).

4.10. Estimation of effects within partnered and single women: Bodily dominance and strength/formidability

We also estimated lower-order interactions and main effects for partnered and single women separately, when Bodily Dominance and Strength/Formidability were entered as male features. The ln(E/P) × Bodily Dominance and ln(E/P) × Strength/Formidability interactions ran strongly in a negative direction for single women. They ran in positive directions for partnered women, though they fell short of significant (The ln(P) × Strength/Formidability was significant for partnered women.) See Tables S16 and S17.

4.11. Summary of hormone × male feature × relationship status effects

In total, we conducted many analyses examining hormone × male feature × Relationship Status effects: ones based on our full model; models removing terms with T; models with grand-mean centered hormone levels; models using residuals on male feature after BMI had been partialled out; models with male age included; models without between-woman hormone terms; models substituting an alternative measure of male feature (Strength/Muscularity/Height, Bodily Dominance, Strength/Formidability) for our Strength/Muscularity

Table 8
Summary results of multilevel regression analyses: Hormone level × strength/muscularity × relationship status interaction effects.

Full Model			T removed			GM centered E/P[b]			With residual S/M		
γ/SE	*t*	*p*	γ/SE	*t*	*p*	γ/SE	*t*	*p*	γ/SE	*t*	*p*
Hormonal predictor: ln(E/P)											
Primary models (from Table 3, main text)											
.05/.02	2.47	**0.014**	.05/.02	2.34	**0.019**	.06/.02	2.78	**0.005**	.04/.02	2.65	**0.008**
Models without between-woman hormone terms (from Table S5)											
.05/.02	2.47	**0.013**	.05/.02	2.34	**0.019**						
Models controlling for male age main effect and interactions (from Table S7)											
.05/.02	2.51	**0.012**	.05/.02	2.37	**0.018**	.06/.02	2.82	**0.005**	.04/.02	2.69	**0.007**
Models without random slope terms											
.05/.02	2.36	**0.018**	.05/.02	2.28	**0.023**	.06/.02	2.63	**0.008**	.04/.02	2.63	**0.008**
Models replacing male strength/muscularity composite with strength/muscularity factor scores (from Table S8)											
.06/.02	2.62	**0.009**	.06/.02	2.47	**0.014**	.07/.03	2.88	**0.004**	.04/.02	2.75	**0.006**
Models replacing male strength/muscularity composite with strength/muscularity/height factor scores (from Table S9)											
.05/.02	2.21	**0.027**	.05/.02	2.08	**0.037**	.06/.02	2.66	**0.008**	.04/.02	2.52	**0.012**
Models replacing male strength/muscularity composite with bodily dominance ratings (from Table 6, Table S14)											
.05/.02	3.28	**0.013**	−.12/.04	3.15	**0.002**	.06/.02	3.25	**0.001**	.05/.02	3.14	**0.002**
Models replacing male strength/muscularity composite with strength/formidability measure (from Table 6, Table S15)											
.07/.02	3.39	**0.001**	.06/.02	3.24	**0.001**	.08/.02	3.54	<**.001**	.05/.02	3.41	**0.001**
Hormonal predictors: estradiol and progesterone entered separately											
Ln(E) and ln(P) entered as hormonal predictors (from Tables 4, Table S6)											
E: .01/.02	0.37		.01/.02	0.31		.01/.03	0.38		.00/.02	0.23	
P:-.05/.02	−2.43	**0.015**	−.05/.02	−2.34	**0.019**	−.06/.02	−2.75	**0.006**	−.04/.02	−2.74	**0.006**
Raw levels of E and P entered as hormonal predictors (from Table S10)											
E: .01/.02	−0.53		−.01/.02	−0.66		−.02/.03	−0.61		−.01/.02	0.31	
P:-.05/.02	−2.30	**0.021**	−.05/.02	2.32	**0.021**	−.05/.02	−2.29	**0.022**	-.04/.02	−2.36	**0.018**

Notes. Ln(E/P) = ln(E) - ln(P). Effects are hence an function of and additive linear composite of ln(E) and ln(P). All quantitative predictors z-scored. Relationship status effect coded: single = −.5, partnered = .5. Observations cross-classified by female raters (*N* = 157), male targets (*N* = 80), and their interaction. Random intercepts for all are modeled. Random slopes, across women, modeled for BMI, Strength/Muscularity, and within-woman hormone measures, except where noted. Inclusion of random slope interactions and covariances selected through model Bayesian Information Criterion fit statistic. Random components and fit statistics reported in Table S2, SOM. *P*-values < .05 bolded. Confidence intervals are not explicitly reported. However, they can be calculated with γ ± 2 × SE.
In the Full Model and T-removed model, hormone levels are centered within-woman. For the GM hormones and With residual S/M models, hormone levels are grand-mean centered. For the Model with residual S/M scores, the male feature (e.g., Strength/Muscularity) is regressed on BMI to remove confounding with BMI.

Table 9
Results of multilevel regression analyses: predictors of sexual attractiveness with cycle phase.

	γ/SE	*t*	*p*
BMI	−1.10/.25	−4.39	<**.001**
Strength/Muscularity (S/M)	1.00/.29	3.49	<**.001**
Relationship status	.20/06	−3.54	<**.001**
Cycle Phase	.07/.04	2.09	**0.037**
Phase × Relationship status	.12/.06	1.95	*0.051*
BMI × Relationship status	−.03/.05	−0.61	
BMI × Phase	−.02/.02	−0.28	
S/M × Relationship status	.03/.05	0.60	
S/M × Phase	.00/.02	0.18	
Rel stat × BMI × Phase	−.02/.04	−0.57	
Rel stat × S/M × Phase	.07/.05	1.59	0.111

Notes. All quantitative predictors z-scored. Relationship status effect coded: single = −0.5, partnered = 0.5. Phase effect codes: −0.5 = luteal; 0.5 = peri-ovulatory. Observations cross-classified by female raters (*N* = 157), male targets (*N* = 80), and their interaction. Random intercepts for all are modeled. Random slopes, across women, modeled for BMI, Strength/Muscularity, and within-woman hormone measures. Inclusion of random slope interactions and covariances selected through model Bayesian Information Criterion fit statistic. Random components and fit statistics reported in Table S24 of SOM. See text and SOM for additional discussion and models. Confidence intervals are not explicitly reported. However, they can be calculated with γ ± 2 × SE.

composite); models in which ln(E) and ln(P) were substituted for ln(E/P); and so on. We present a summary of the hormone × male feature × Relationship Status effects emerging from these analyses in Table 8. As can be seen, the effect robustly emerges across analyses.

4.12. Using cycle phase as a predictor

In secondary analyses (Gangestad, Dinh, et al., 2018), we substituted cycle phase for ln(E/P). The Cycle Phase × Strength/Muscularity × Relationship Status interaction falls short of statistical significance, *t* = 1.59, *p* = .111. See Table 9. The contrast between this result and the comparable ln(E/P) 3-way interaction requires an explanation. If hormones drive cycle shifts, hormonal associations should exceed cycle phase associations. Some phases may be mischaracterized, and some cycles anovulatory. In Roney and Simmons' (2013) sample, 33% of all cycles were anovulatory or evidenced luteal insufficiency, judged by small progesterone rises. Some of these cases surely exist in Jünger et al.'s sample. An LH surge (especially one detectable with the very high sensitivity strips Jünger et al. used) is not necessarily indicative of ovulation; in anovulatory cycles, LH may rise, though surges may be blunted (e.g., Wu & Cowchock, 1983). Lynch et al. (2014) found that, among cycles classified as anovulatory based on failure to cross a threshold of luteal progesterone level (akin to that used by Roney & Simmons, 2013), the LH increase from baseline still achieved 70% of the increase in cycles classified as ovulatory—levels very likely detectable with Jünger et al.'s high sensitivity method. Perhaps even more importantly, and as already noted (see fn 7), Jünger et al. conducted 14% of fertile phase sessions 2+ days after an LH surge; the majority of these sessions would be during the luteal phase and non-conceptive. (Wetzels & Hoogland, 1982 found that the initial LH surge, measured in serum, occurred 11–24 h prior to ovulation, as detected by ultrasonography. Conception risk drops steeply after ovulation.) Another 12% were conducted one day after the LH surge; a portion of these would likely also have been during non-conceptive occasions (e.g., Dunson, Baird, Wilcox, & Weinberg, 1999) (see fn 7). The timing of high fertility sessions, relative to the LH peak, varied by up to 8 days (3 days prior to a surge to 4 days after). Hence, Jünger et al.'s measure of "phase", even among cycles with positive LH surges, possesses a considerable degree of noise. Estradiol and progesterone levels, by contrast, were time-locked with session and, hence, concurrent with assessments of preferences.

Progesterone levels during truly conceptive peri-ovulatory and mid-luteal phases should overlap little (Ellison, 1993). Thus, in exploratory analyses, we restricted cases to those exhibiting no or limited overlap through a range of procedures. The Cycle Phase × Strength/Muscularity × Relationship Status interactions were significant in these subsets. Analyses are reported in Table S23. We fully acknowledge and emphasize that these analyses add very little, if any, *independent* evidence for cycle effects beyond what hormonal associations offer. If ln(E/P) and progesterone levels interact with relationship status to affect preferences, the interaction effect of phase and relationship status on preferences will increase when cases are selected to accentuate progesterone levels between fertile and non-fertile sessions—in effect, potentially removing luteal-phase cases misclassified as being within the fertile-phase, as well as luteal-phase cases with progesterone levels reflective of non-conceptive cycles. These findings, then, merely illustrate implications of analyses already presented; in no way do they constitute a novel empirical test. That said, these implications are not trivial. If steroid hormones regulate cycle shifts, then hormonal measures should produce larger effects than cycle phase, especially when cycle phase is a noisy measure. Null findings with respect to phase should not be used to infer the null hypothesis. The hormonal associations we find invite an alternative explanation for weaker findings for phase: Jünger et al.'s measure of phase does not tap the drivers of cycle shifts as well as direct hormonal measures do.

5. Contrasting results

5.1. Null conclusions and main effects of hormones on general attraction?

Jünger et al. presented preregistered analyses examining whether women's cycle phase and ovarian hormones moderate women's sexual attraction to men's muscular features. They found no evidence for such effects, "*indicating no shifts in preferences for specific traits*" (p. 419); cycle shifts "*do not seem to alter preferences for body characteristics at all, leaving no room for cycle shifts in mate preferences for masculine characteristics or any other assumed indicators of good genes*" (p. 421; emphasis added).

By contrast, our analyses on Jünger et al.'s data yields suggestive evidence that a measure of men's Strength/Muscularity (controlling for BMI) more strongly predicts partnered women's sexual attraction when estradiol levels are high relative to their progesterone levels. Single women exhibit an opposite pattern. Analyses using a measure of male bodies' formidability or a global rating of bodily dominance yield similar hormonal moderation effects. These key results are robust to inclusion/exclusion of control variables (age, women's testosterone) and exclusion/inclusion of outliers. The patterns suggested by these analyses contrast with Jünger et al.'s conclusions: Women's hormone levels, in concert with their relationship status, moderate associations of men's muscular features with women's sexual attraction. When women in relationships produce concentrations of ovarian hormones characteristic of high conception risk, they may be especially sexually attracted to strong, muscular men (independent of BMI); single women may show opposite associations. These patterns are driven by women's progesterone levels. As well, these analyses provide evidence that romantically involved women with a hormonal profile of high conception risk may be especially attracted to bodies that are relatively lean—bodies of low BMI, with measures of muscularity controlled.

Jünger et al. claim that, when conceptive, partnered women rate men's bodies in general as more attractive. We find more mixed effects using hormonal predictors (with $p > .05$ in most analyses). These effects may be real, but they may also be qualified by relationship status and male features. Among partnered women, ln(E/P) may be associated with sexual attraction to men scoring high on Strength/Muscularity but *not* (or minimally) with sexual attraction to men scoring low on Strength/Muscularity.

We fully acknowledge that, though relationship status-hormone interaction effects appear to be robust across analyses, simple effects for partnered and single women separately do not consistently yield significant effects. Across 4 measures—Strength/Muscularity, Strength/Muscularity/Height, Bodily Dominance, and Strength/Formidability—and 2 hormonal measures—ln(E/P) and ln(P)—50% (4/8) of analyses yielded $p < .05$ for hormonal effects on partnered women's preferences; 62% (5/8) yielded $p < .05$ for hormonal effects on single women's preferences. No definitive conclusions in this regard can hence be reached. But just as results do not yield definitive evidence for significant hormonal moderation for partnered or single women, they surely too do not yield evidence of no effects, contrary to Jünger et al.'s conclusions (e.g., Amrhein, Greenland, & McShane, 2019).

5.2. What explains the differences?

Our analyses find support for hormonal effects on preferences. Jünger et al.'s did not. What factors made the difference? We focus on three mentioned previously, along with one other.

5.2.1. Examining the moderating role of relationship status

We start with the obvious: We examined effects—hormone × male feature × relationship status interactions—that Jünger et al. did not, despite preregistering a hypothesis directly pertaining to these effects.

5.2.2. Controlling for preference for BMI

Jünger et al. only controlled for the main effect of BMI. Failing to control for BMI interactions as well leaves confounds in preference shifts. When we too entered *only* BMI's main effect, the critical ln(E/P) × Strength/Muscularity × Relationship Status effect (initial analysis, Table 4) weakened, $t = 2.25$, $p = .025$.

5.2.3. Compositing features vs. pitting them against one another

In primary analyses, Jünger et al. entered male features simultaneously. Tests on each can detect unique effects only, weakening power to detect shared effects. When we similarly entered Strength and Upper Arm Circumference simultaneously, neither ln(E/P) × male feature × Relationship Status interaction effect was significant: $t = 1.50$, $p = .133$; $t = 1.48$, $p = .138$, respectively. With BMI interactions also uncontrolled—as in Jünger et al.'s analyses—effects were weaker yet: $t = 1.42$, $p = .156$; $t = .86$, $p = .67$. Jünger et al.'s primary analytic approach was not especially sensitive to detecting hypothesized effects.

5.2.4. Random slope effects

We add one feature. We modeled random slope effects for BMI, male features, hormones, and phase across women. That is, our models estimated variation across women in sensitivity of ratings to male features and hormones. Random slope effects were generally very large, estimates often 5+ times their standard errors; their inclusion greatly increased model fit (see S24). That may well be because the standard deviation of individual women's ratings differed substantially: from <1 to >4 (i.e., women used different ranges of the scale). Jünger et al. did not model these random slopes. Yet exclusion of meaningful random slope terms can greatly overestimate the robustness of some fixed effects, largely because error terms are underestimated (e.g., Judd, Westfall, & Kenny, 2012; Barr, Levy, Scheepers, & Tily, 2013).

Jünger et al.'s results most affected by inclusion of random slopes pertain to their primary positive take-homes. They report robust Cycle Phase and Cycle Phase × Relationship Status effects on sexual attraction.,." When we repeated Jünger et al.'s analysis including a random slope component, fit improved substantially: BIC change = −306.1. (See S24. BIC difference > 10 is typically considered large; e.g., Vrieze, 2012.). While the Cycle Phase main effect remained significant, it was less impressive: $t = 2.09$, $p = .037$. The relationship status interaction fell short of being significant, $p = .051$. See Table 9. In our analyses that used within-woman or grand-mean centered ln(E/P) rather than cycle phase, ln(E/P) never interacted with relationship status to predict sexual attraction. See Table 4.

5.2.5. *Log-transformation*

In our planned analyses, we entered log-transformed hormone levels, following common practice within endocrinological research. In Table S10, we present analyses that examined preferences using untransformed estradiol and progesterone levels. As we would anticipate (see Footnote 7; see also Footnote 8), the untransformed progesterone × Strength/Muscularity × Relationship Status interaction was slightly weaker than the ln(P) × Strength/Muscularity × Relationship Status interaction, though not markedly so.

5.3. *Correlation between mean ratings across sessions*

We address one additional argument Jünger et al. made. They emphasized that there is "*no room* for differential effects of masculinity cues" (p. 417; emphasis added) because the rank order correlation of sexual attractiveness ratings across men for high and low conception risk women is nearly perfect (Spearman rank $\rho = .998$). This argument misconstrues the impacts of differential effects. When some women weight an influential feature more than others do, rank ordering across women need not be greatly affected. On that particular feature, men have a fixed rank-ordering. Weighting the feature more, all else equal, will increase the *dispersion* of ratings as a function of the feature (i.e., increase the regression slope), but the ordering of how ratings of men are affected by the feature *remains unchanged*.[16] Ordering of men on that feature may differ from ordering on other features, such that differential weighting will shift overall, weighted ordering somewhat. But changes may be minimal. To demonstrate this, we analyzed mean ratings given to men by women at high and low ln(E/P). The regression weights of Strength/Muscularity and BMI were greater for mean ratings at high ln(E/P), yet the two sets of ratings correlated .993; see S25 in SOM for details. Contrary to Jünger et al.'s claims, a near-perfect correlation does *not* entail that there is "no room" for differential effects.

5.4. *Effect size estimation*

Statistically significant effects may be inconsistent with the null hypotheses, while nevertheless reflecting effect sizes that are inconsequential. Are the effects we report theoretically meaningful? Within partnered women, the per unit impact of Strength/Muscularity on attractiveness ratings is estimated to be 8% greater when ln(P) is 1 s below the mean (21st percentile) compared to when ln(P) is 1 s above the mean (75th percentile; (.879 + .0326)/(.879–.0326); Table 5). This difference in impact produces a 16% boost in variance in attractiveness ratings of women 1 s below mean ln(P) associated with Strength/Muscularity relative to ratings of women +1 s above mean ln(P) ($1.08^2 = 1.16$). For women at extremes on ln(P), the 5th and 95th percentiles ($-1.32\,s$ and $1.55\,s$ from the mean, respectively), this difference in variance is naturally larger, 24%. Differences are of similar size for single women, but in the opposite direction. Differences in impact strike us as potentially meaningful. At the same time, a 95% confidence interval around effect sizes includes ones both near-zero and very substantial – double the point estimate (variance differences of 33% and 51% for the two comparisons above). The current data do not allow one to pinpoint effect sizes with sufficient precision to judge their theoretical meaningfulness or practical impact.

Jünger et al. repeatedly presented women with headless digital figures lacking some human-typical features, such as realistic skin tone. In so doing, they enhanced experimental control by stripping out individuating features aside from bodily shape, but likely at a cost of ecological validity and psychological realism. Women do not encounter, evaluate, or respond to such male figures in everyday life. Of course, they may evaluate their attractiveness, in certain regards, using processes designed to evaluate "real" male bodies. But one cannot assume that effect sizes revealed in Jünger et al.'s study directly generalize to effect sizes in women's evaluations of real bodies. This point is not a criticism of Jünger et al.'s study; the trade-off between control and realism entailed by their study design is very reasonable. At the same time, this trade-off implies that an estimated effect size need not match effect sizes in women's everyday life. We stress that additional work is needed to fully assess the meaningfulness of effects in ecological conditions.

5.5. *Interpretation*

What evolutionary account explains hormonal moderation of preferences for muscularity? Do these data yield evidence for the good genes interpretation of hormonal effects? Though the evidence we present could potentially be consistent with a good genes framework, more work is needed to clarify appropriate interpretation. Several key aspects of the findings must be addressed.

First, no preference shift independent of relationship status emerged; only romantically involved women displayed the preference shifts predicted by the good genes account. As Jünger et al. note, particular forms of the good genes hypothesis (such as the dual mating hypothesis; Pillsworth & Haselton, 2006) expect moderation by relationship status. But other possible explanations for this moderation should also be considered, including Type I error, conjectures that non-conceptive sex plays special roles in partnered women (Grebe, Gangestad, Garver-Apgar, & Thornhill, 2013), and other perspectives on human mating (Emery Thompson & Muller, 2016).

Second, the 3-way interaction is not a simple attenuated 2-way interaction. Based on good genes thinking, one might expect a large positive ln(E/P) × muscularity interaction for women in relationships and a small or zero interaction for single women. Yet the 3-way interaction is driven by two 2-way interactions in opposite directions: positive for partnered women and negative for single women. For analyses examining preferences for Bodily Dominance, 2-way interactions were robust for single women but not for partnered women. Sampling variability could of course play a role (perhaps the true interaction *is* an attenuated one), but that possibility begs for additional studies.[17]

Third, changes in romantically involved women's progesterone are associated with changes in mate preferences in this sample. Estradiol-linked changes were generally not suggested. Yet other studies link variation in estradiol to levels of sexual interest (e.g., Grebe et al., 2016; Roney & Simmons, 2013).

[16] Imagine, for instance, that ratings were a function of a single cue, but some women made greater discriminations based on the cue than others. (E.g., some women prefer the cue by a lot, others prefer it by a little.) The correlation between each woman's ratings and the cue would be 1.00, and women's ratings would correlate with each other 1.00. Differential use of the cue across women would be reflected in variances, with women making stronger discriminations based on the cue giving more variable ratings.

[17] One reason to be cautious about drawing conclusions concerning the relative 2-way hormone × male feature interactions for single and partnered women is that they vary across measures of male feature. Hence, though the 2-way interaction is stronger for single women using Bodily Dominance as a measure, it is stronger for partnered women when Strength/Muscularity/Height is used. Again, more data are needed.

5.6. An independent demonstration

Since we conducted these analyses, we learned of another, recently published study that found a similar interaction. Marcinkowska, Kaminski, Little, and Jasienska (2018) examined preferences for male bodily masculinity in a sample of 102 women. Their preference measure consisted of just 3 items and possessed low internal consistency. Furthermore, sample size was smaller than Jünger et al.'s; in light of reduced power, results must be interpreted cautiously. Marcinkowska et al. reported, however, a significant within-woman Progesterone × Relationship Status effect on preferences, running in the same direction as we report here. We note that, unlike in our analyses, the simple effect of progesterone for partnered women was not significant (and, indeed, was near-zero). The simple effect for single women ran in a positive direction. Though these results give additional reason to think that the interaction effect we report is robust, better estimation of simple effects for partnered and single women requires more research.[18]

6. Reflections on preregistration and related issues

Preregistration of analyses is a valued methodological quality that we endorse. That said, it is not the sole or most important one. First and foremost, a set of analyses should appropriately assess a conceptual question, which preregistration itself does not ensure; as illustrated by the current dataset, two different analyses yield contrasting conclusions. One need not decide which analyses best address major issues to appreciate the illustration. As discussed elsewhere (e.g., PsychMAP, 2018), consumers may heuristically use preregistration as a cue that the authors of a study have selected the "best" analytical strategy, yet doing so entails risk.

We offer here several reflections on preregistration and related issues.

Robustness. Preregistration constrains which analyses are "confirmatory." Much responsibility, then, is placed on researchers to carefully think through analyses prior to preregistration. Even ardent proponents of preregistration can admit that preregistered analyses that inadequately address key conceptual questions may deter, not facilitate, proper understanding. Sometimes, authors cannot fully anticipate which analyses appropriately address a set of questions. Best analyses may hinge on features of the data (presently, illustrated by validation of muscular features). And rather than foreseeing a single best strategy, researchers may envision a set of analyses across which robustness may be judged. Preregistration may encourage authors to capture their preplanned hypothesis testing in a single analysis, thereby downplaying a role for validity and robustness checks.

6.1. Robustness applies to null results too

Scholars appreciate robustness as a quality of positive results (e.g., Arslan et al., in press); indeed, Jünger et al. analyzed their data in a variety of ways. Yet it is desirable for null results too. After all, null conclusions reflect absence of evidence for effects, yet null results are often interpreted as evidence of absent effects. To justify the latter, the former cannot be thin. Presently, Jünger et al. found no interactions between cycle phase and individual male features. Yet they did not examine hormonal associations—a priori, analyses that should have greater power than the ones they conducted—or moderation by relationship status. Still, they concluded that their findings indicate "*no shifts in preferences for specific traits*"—an explicit claim of *evidence for absence*, not absence of evidence (see also Amrhein et al., 2019).

6.2. Preregistration and up-down thinking in hypothesis-testing

As argued by others (e.g., Amrhein et al., 2019; Cumming, 2014), hypothesis-testing cultivates simple up-down thinking: An alternative hypothesis is supported or not, favoring a null hypothesis. A certain use of preregistered studies may inadvertently reinforce this thinking. In its ideal form, a straightforward preregistered test is performed, yielding evidence for an alternative hypothesis or not. If not, that is it; additional analyses, not being "confirmatory," are non-informative with respect to hypothesis-testing and are thereby implicitly discouraged.[19] This thinking is illustrated by Jünger et al.'s null conclusions based on particular null findings, as are its risks.

Naturally, Type I and Type II errors trade off. If Type I errors are especially aversive, additional Type II errors could be warranted. But this reasoning itself assumes simple up-down thinking. In fact, scientific inference should not be so simplistic. Evidence typically permits only degrees of scientific belief (whether in probability [e.g., Salmon, 1970; Carnap, 1947] or truth-likeness [Popper, 1963] terms), a point that applies to individual studies. In conjunction with past findings, it informs belief updating (explicitly Bayesian or not); only rarely will it justify definitive up-down answers. Those alarmed by the replication crisis rightly deem simplistic hypothesis-testing a bad actor. Through publication bias, *p*-hacking, post-hoc hypothesizing, overinterpretation of findings, and non-transparency, it inflates Type I errors. The solution, however, should not be similarly simplistic thinking, where Type II errors substitute for Type I errors. Rather, cautious and nuanced discussion of what findings mean—less definitive and more modest than what simple up-down thinking invites—should be fostered (Amrhein et al., 2019).

Because it invites simple binary, up-down thinking, Amrhein et al. (2019) propose that the concept of statistical significance be abandoned altogether (though, we stress, they do not argue that *p*-values are meaningless and useless). Along similar lines, in a recent commentary Gelman (2018) recommended that "we should stop labeling replications as successes or failures and instead use continuous measures to compare different studies" (p. xxx). Binary labels "get us into trouble with their implication that there is some criterion under which a replication can be said to succeed or fail. Do we just check whether p < .05? That would be a very noisy rule…" (p. xxx). A focus on effect size estimation through aggregation of data over time dispenses with the idea of Type I and Type II errors altogether (though it recognizes potential errors in effect size estimation; Cumming, 2014; Gelman & Carlin, 2014).

6.3. Exploration and the total evidence rule

Preregistered, confirmatory analysis is often pitted against exploratory analysis, when, in fact, the two are complementary (e.g., Jebb, Parrigon, & Woo, 2017). Preregistered analyses address targeted

[18] Both Marcinkowska et al. (2018) and Debruine et al. (2019) also report robust between-woman (i.e., woman-mean) Progesterone × Relationship Status interactions predicting women's preferences for facial masculinity. These interactions run in the same direction as we and Marcinkowska et al. find for within-woman Progesterone × Relationship Status: in a positive direction for single women and a negative direction for partnered women. Debruine et al. (2019) argue that, because they and Marcinkowska et al. (2018) found no within-woman Progesterone × Relationship Status interactions predicting facial masculinity preferences, the between-woman Progesterone interactions likely do not reflect direct effects of progesterone. That said, we caution against interpreting a non-significant effect as evidence of "no effect" (e.g., Amrhein et al., 2019). The issue of whether these interactions are related and due to direct effects of progesterone is, in our view, not yet fully resolved.

[19] Interestingly, from a Bayesian perspective one can argue that the distinction between planned versus post-hoc tests is not a substantive one, and thus is not the main point of preregistration (e.g., Dienes, 2016). While the distinction has its uses, it should be employed critically while being aware of its scope and limitations.

questions. Exploratory analyses permit understanding of data in ways unanticipated (e.g., contingent on unexpected results), and may suggest directions for future theory development and empirical investigation. Furthermore, they permit examinations of robustness not anticipated during preregistration. Though commonly referred to as "exploratory" because they were not explicitly preplanned, these examinations may readily be at least as grounded in pertinent theory and pertinent bodies of evidence as planned analyses. Carnap (1947) argued that, when applying inductive logic to estimate the probability of an event, one should consider the full totality of evidence pertinent to the induction. Though philosophers have debated the foundations of the "total evidence" principle (e.g., Suppes, 1966), it captures an idea most scientists endorse: In evaluating the strength of evidence for an interpretation, one should not ignore any important information pertinent to evaluating the interpretation. Unwittingly, however, sharp demarcations between confirmatory and exploratory analysis, in conjunction with simple up-down inferential thinking, may encourage violations—especially regarding null conclusions. Surely, many analyses Jünger et al. did not conduct are still pertinent to their null conclusions: e.g., hormonal associations; moderation by relationship status; analyses on Bodily Dominance ratings. Hence, their null conclusions ignored important components of the "total evidence" contained in their own data. We are wary of practices that encourage these outcomes.

6.4. Broader costs of null conclusions

Individual effects in single studies are rarely empirically isolated phenomena. Rather, they fit into, and hence speak to, larger conceptual networks (e.g., Fiedler, Kutzner, & Krueger, 2012). Here, hormone-associated shifts speak to broader, integrative theories within evolutionary psychology. Jünger et al. emphasize this point; they draw theoretical implications of their results, arguing that null conclusions weigh against good genes accounts and in favor of motivational priorities perspectives on cycle shifts. These arguments could affect the fate of future research paths taken and foregone; researchers generally avoid testing theories that are (rightly or wrongly) perceived as "dead." However, integrative ideas with heuristic potential are not easy to come by. There is value to "pulling weeds," that is, discarding false claims. At the same time, premature assertions of the null—especially if bolstered by the aura of a preregistered study—can mistakenly "pull" generative stocks, the costs of which can be substantial. One can hence argue that, *even if most novel integrative ideas are wrong*, on balance premature null conclusions deter scientific progress (e.g., Fiedler, 2017; Fiedler et al., 2012). Naturally, this point is a general one, not specific to the current theoretical context.

To conclude, it is worth stressing that our analyses are not proof that preference shifts exist. Jünger et al.'s conclusions may yet be right. At the same time, Jünger et al.'s data do not constitute solid evidence for a null conclusion. Our analyses provide reason to think that relationship status moderates shifts in preferences for muscularity, and suggest new hypotheses about preferences for leanness (which, in conjunction with muscularity, may reflect physical fitness) and shifts among single women. Naturally, more data are needed to address these matters. These conclusions may be modest, and—we think—appropriately so. Though motivated by good intentions, some thinking behind preregistration, and the deep concerns about non-replicability that drive it, may not encourage such modesty. Rather, for reasons we discuss above, it may inadvertantly foster the approach that led Jünger et al. to prematurely draw null conclusions in this particular case.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.evolhumbehav.2019.05.005.

References

Alvarado, L. C., Muller, M. N., Thompson, M. E., Klimek, M., Nenko, I., & Jasienska, G. (2016, March). Men's reproductive ecology and diminished hormonal regulation of skeletal muscle phenotype: An analysis of between-and within-individual variation among rural Polish men. *American Journal of Physical Anthropology*. *Vol. 159*Hoboken NJ: Wiley-Blackwell 78–78.

Amrhein, V., Greenland, S., & McShane, B. (2019). Comment: Retire statistical significance. *Nature, 567*, 305–307.

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 behaviour. Journal of Personality and Social Psychology DOI: https://doi.org/10.1037/pspp0000208.

Baird, D. D., Weinberg, C. R., Wilcox, A. J., & McConnaughey, D. R. (1991). Using the ratio of estrogen and progesterone metabolites to estimate the day of ovulation. *Statistics in Medicine, 10*, 255–266.

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

Carnap, R. (1947). *Meaning and necessity*. Chicago, IL: University of Chicago Press.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29.

Debruine, L. M., Hahn, A. C., & Jones, B. C. (2019). Does the interaction between partnership status and average progesterone level predict women's preferences for facial masculinity? *Hormones and Behavior, 107*, 80–82.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72*, 78–89.

Dinh, T., Pinsof, D., Gangestad, S. W., & Haselton, M. G. (2017). Cycling on the fast track: Ovulatory shifts in sexual motivation as a proximate mechanism for regulating life history strategies. *Evolution and Human Behavior, 38*, 685–694.

Dixson, A. (2013). *Primate sexuality: Comparative studies of the prosimians, monkeys, apes, and humans*. Oxford University Press.

Dunson, D. B., Baird, D. D., Wilcox, A. J., & Weinberg, C. R. (1999). Day-specific probabilities of pregnancy based on two studies with imperfect measures of ovulation. *Human Reproduction, 14*, 1835–1839.

Ellison, P. T. (1993). Measurements of salivary progesterone. *Annals of the New York Academy of Sciences, 694*, 161–176.

Ellison, P. T. (2003). Energetics and reproductive effort. *American Journal of Human Biology, 15*(3), 342–351.

Emery Thompson, M., & Muller, M. N. (2016). Comparative perspectives on human reproductive behavior. *Current Opinion in Psychology, 7*, 61–66.

Fessler, D. M. T., Holbrook, C., & Snyder, J. K. (2012). Weapons make the man (larger): Formidability is represented as size and strength in humans. *PLOS ONE, 7*(4), https://doi.org/10.1371/journal.pone.0032751.

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science, 12*, 46–61.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*, 661–669.

Frederick, D. A., & Haselton, M. G. (2007). Why is muscularity sexy? Tests of the fitness-indicator hypothesis. *Personality and Social Psychology Bulletin, 33*, 1167–1183.

Gangestad, S. W., Dinh, T., Grebe, N. M., Gildersleeve, K., Emery Thompson, M., & Haselton, M. G. (2018). A replication study examining effects of cycle phase and hormonal indicators on two female mate preferences. *Preregistration posted on Open Science Framework*. https://osf.io/4x7ub/?view_only=3651613e41ea4e0c8d8abc97cc6cfc3c.

Gangestad, S. W., Grebe, N. M., Gildersleeve, K., & Haselton, M. G. (2018). *Are ovulatory shifts in women's mate preferences robust? Selection models say it depends.* (Manuscript under revision).

Gelman, A. (2018). Don't recognize replications as successes or failures. *Behavioral and Brain Sciences, 41*. https://doi.org/10.1017/S0140525X18000638, e128.

Gelman, A., & Carlin, J. B. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*, 641–651.

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. URL http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

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*, 1205–1259.

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

Grebe, N. M., Emery Thompson, M., & Gangestad, S. W. (2016). Hormonal predictors of women's in-pair and extra-pair sexual attraction in natural cycles: Implications for extended sexuality. *Hormones and Behavior, 78*, 211–219.

Grebe, N. M., Gangestad, S. W., Garver-Apgar, C. E., & Thornhill, R. (2013). Women's luteal-phase sexual proceptivity and the functions of extended sexuality. *Psychological Science, 24*, 2106–2110.

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, 14*, 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*, 256–259.

Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analyses as a foundation of inductive research. *Human Resource Management Review, 27*, 265–276.

Jones, B. C., Hahn, A. C., Fisher, C., Wang, H., Kandrik, M., & DeBruine, L. M. (2018a). General sexual desire, but not desire for uncommitted sexual relationships, tracks changes in women's hormonal status. *Psychoneuroendocrinology, 88*, 153–157.

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

Jones, K. A. (1996). Summation of basic endocrine data. In G. A. Gass, & H. M. Kaplan (Vol. Eds.), *Handbook of Endocrinology*(2nd ed.). *Vol. 1. Handbook of Endocrinology* (pp. 2–42). Boca Raton FL: CRC Press.

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

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*, 412–423. https://doi.org/10.1016/j.evolhumbehav.2018.03.007.

Jünger, J., & Penke, L. (2016). The effects of ovulatory cycle shifts in steroid hormones on female mate preferences for body masculinity, voice masculinity and social dominant behavior. Preregistration. *Open Science Framework*. https://osf.io/u3y7a/.

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.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*, 1–21.

Kutner, M. A., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models* (4th ed.). New York: McGraw-Hill/Irwin.

Lindsay, C. S. (2017). Editorial: Sharing data and materials in *Psychological Science*. *Psychological Science, 28*, 699–702.

Lipson, S. F., & Ellison, P. T. (1996). Comparison of salivary steroid profiles in naturally occurring conception and non-conception cycles. *Human Reproduction, 11*, 2090–2096.

Lukaszewski, A. W., Simmons, Z. L., Anderson, C., & Roney, J. R. (2016). The role of physical formidability in human social status allocation. *Journal of Personality and Social Psychology, 110*(3), 385–406. https://doi.org/10.1037/pspi0000042.

Lynch, K. E., Mumford, S. L., Schliep, K. C., Whitcomb, B. W., Zarek, S. M., Pollack, A. Z., et al. (2014). Assessment of anovulation in eumenorrheic women: Comparison of ovulation detection algorithms. *Fertility and Sterility, 102*, 511–518.

Marcinkowska, U. M., Kaminski, G., Little, A. C., & Jasienska, G. (2018). Average ovarian hormone levels, rather than daily values and their fluctuations, are related to facial preferences among women. *Hormones and Behavior, 102*, 114–119.

Millar, M. (2013). Menstrual cycle changes in mate preferences for cues associated with genetic quality: The moderating role of mate value. *Evolutionary Psychology, 11*, 18–35.

Nestler, S., & Back, M. D. (2013). Applications and extensions of the lens model to understand interpersonal judgments at zero acquaintance. *Current Directions in Psychological Science, 22*, 374–379.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*. https://doi.org/10.1126/science.aac4716.

Pillsworth, E. G., & Haselton, M. G. (2006). Women's sexual strategies: The evolution of long-term bonds and extra-pair sex. *Annual Review of Sex Research, 17*, 59–100.

Popper, K. R. (1963). *Conjectures and refutations.* London: Routledge.

PsychMAP (2018). In *Facebook* [group page]. Retrieved May 25, 2018, from https://www.facebook.com/groups/psychmap/permalink/580032225707037/.

Roney, J. R., & Simmons, Z. L. (2013). Hormonal predictors of sexual motivation in natural menstrual cycles. *Hormones and Behavior, 63*, 636–645.

Roney, J. R., & Simmons, Z. L. (2016). Within-cycle fluctuations of progesterone negatively predict changes in both in-pair and extra-pair desire among partnered women. *Hormones and Behavior, 81*, 45–52.

Roney, J. R., & Simmons, Z. L. (2017). Ovarian hormone fluctuations predict within-cycle shifts in women's food intake. *Hormones and Behavior, 90*, 8–14.

Salmon, W. (1970). In R. Stuewer (Ed.). *Historical and philosophical perspectives of science* (pp. 68–86). Minneapolis, MN: University of Minnesota Press Bayes theorem and the history of science.

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Sollberger, S., & Ehlert, U. (2016). How to use and interpret hormone ratios. *Psychoneuroendocrinology, 63*, 285–297.

Suppes, P. (1966). Probabilistic inference and the concept of total evidence. In J. Hintikka, & P. Suppes (Eds.). *Aspects of inductive logic* (pp. 49–65). Amsterdam: North-Holland Publishing Co.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*, 228–243.

Wallen, K. (2013). Women are not as unique as thought by some: Comment on “hormonal predictors of sexual motivation in natural menstrual cycles,” by Roney and Simmons. *Hormones and Behavior, 63*(4), 634–635. https://doi.org/10.1016/j.yhbeh.2013.03.009.

Welling, L. L., Jones, B. C., DeBruine, L. M., Conway, C. A., Smith, M. L., Little, A. C., ... Al-Dujaili, E. A. (2007). Raised salivary testosterone in women is associated with increased attraction to masculine faces. *Hormones and Behavior, 52*, 156–161.

West, S. G., Ryu, E., Kwak, O.-M., & Chan, H. (2011). Multilevel modeling: Current and future applications in personality research. *Journal of Personality, 79*, 1–50.

Wetzels, L. C. G., & Hoogland, H. J. (1982). Relation between ultrasonographic evidence of ovulation and hormonal parameters: Luteinizing hormone surge and progesterone rise. *Fertility and Sterility, 37*, 336–341.

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.

Wu, C. H., & Cowchock, F. S. (1983). Daily blood hormone levels related to the luteinizing hormone surge in anovulatory cycles. *Fertility and Sterility, 39*, 39–43.

Evolution and Human Behavior 40 (2019) 517–525

Contents lists available at ScienceDirect

Evolution and Human Behavior

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

Commentary

No robust evidence for cycle shifts in preferences for men's bodies in a multiverse analysis: A response to Gangestad, Dinh, Grebe, Del Giudice, and Emery Thompson (2019)

Julia Stern[a,*], Ruben C. Arslan[b], Tanja M. Gerlach[a], Lars Penke[a]

[a] Department of Psychology & Leibniz ScienceCampus Primate Cognition, University of Goettingen, Gosslerstrasse 14, 37073 Goettingen, Germany
[b] Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

ARTICLE INFO

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, Thornhill, & Garver-Apgar, 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, 2019, 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 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., criticize their analytical decisions that they claim to have derived from their preregistration, and clarify a potentially misleading reporting of an independent study by Marcinkowska, Kaminski, Little, and Jasienska (2018).

[*] Corresponding author.
E-mail address: julia.stern@psych.uni-goettingen.de (J. Stern).

[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.

https://doi.org/10.1016/j.evolhumbehav.2019.08.005
Received 20 August 2019; Accepted 23 August 2019

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, Kordsmeyer, 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 & Jasienska, 2005), sexual dimorphism in body shape (Little, Jones, & Burriss, 2007), and muscularity (Gangestad, Garver-Apgar, Simpson, & Cousins, 2007), especially in the light of reported null replications (Marcinkowska, Galbarczyk, & Jasienska, 2018; Peters, Simmons, & Rhodes, 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, Dunn, Hopkins, & Kang, 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 analysed 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, Hunt, Puts, Ostner, & Penke, 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 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). 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 $r = 0.06$, 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 ($r = 0.32$, 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-

[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.

[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.

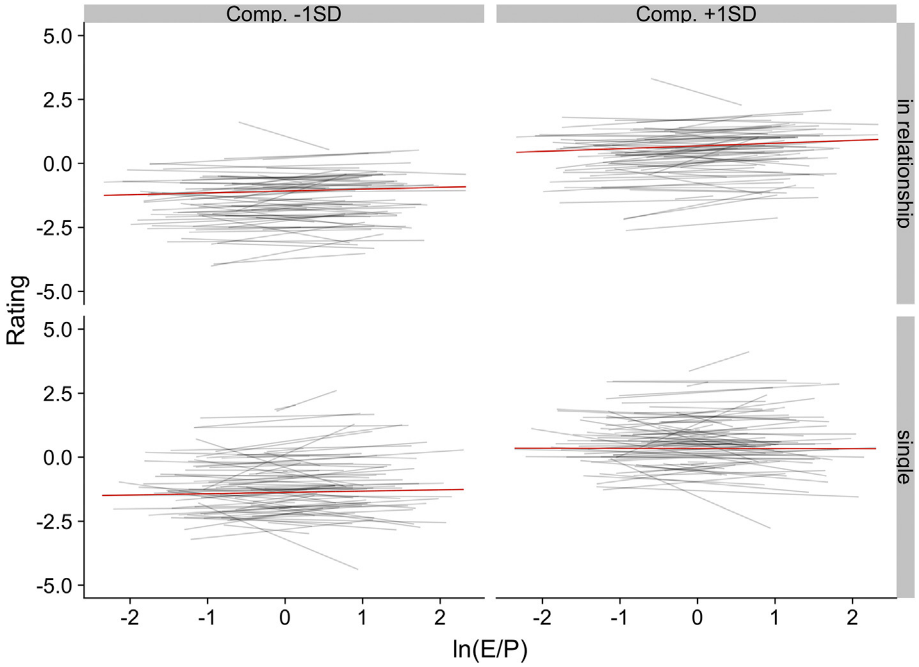

Fig. 1. This spaghetti plot shows that only a very small amount of the variation in slopes between women (gray lines) is explained by the moderators ln E/P and relationship status. For the most part, women consistently prefer men who are higher in muscularity (Gangestad et al.'s S/M component). The slopes 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.

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 Fig. 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, Galbarczyk, & Jasienska, 2018). 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 $ps > 0.24$; see Tables S3 and S4). Again, this fragility of their results was not acknowledged by Gangestad et al.

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 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 Fig. 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.

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 0.94 to detect a two-way interaction effect of d = 0.35. 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, Kaminski, et al. (2018) results

In Section 5.7 of their reanalysis, Gangestad et al. report an effect of Marcinkowska, Kaminski, et al. (2018) study. Here, they state that Marcinkowska, Kaminski, et al. (2018) 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, Galbarczyk, & Jasienska, 2018). Marcinkowska, Kaminski, et al. (2018) 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, Kaminski, et al. (2018) 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.

Fig. 2 shows a garden of forking paths: it showcases the possible

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

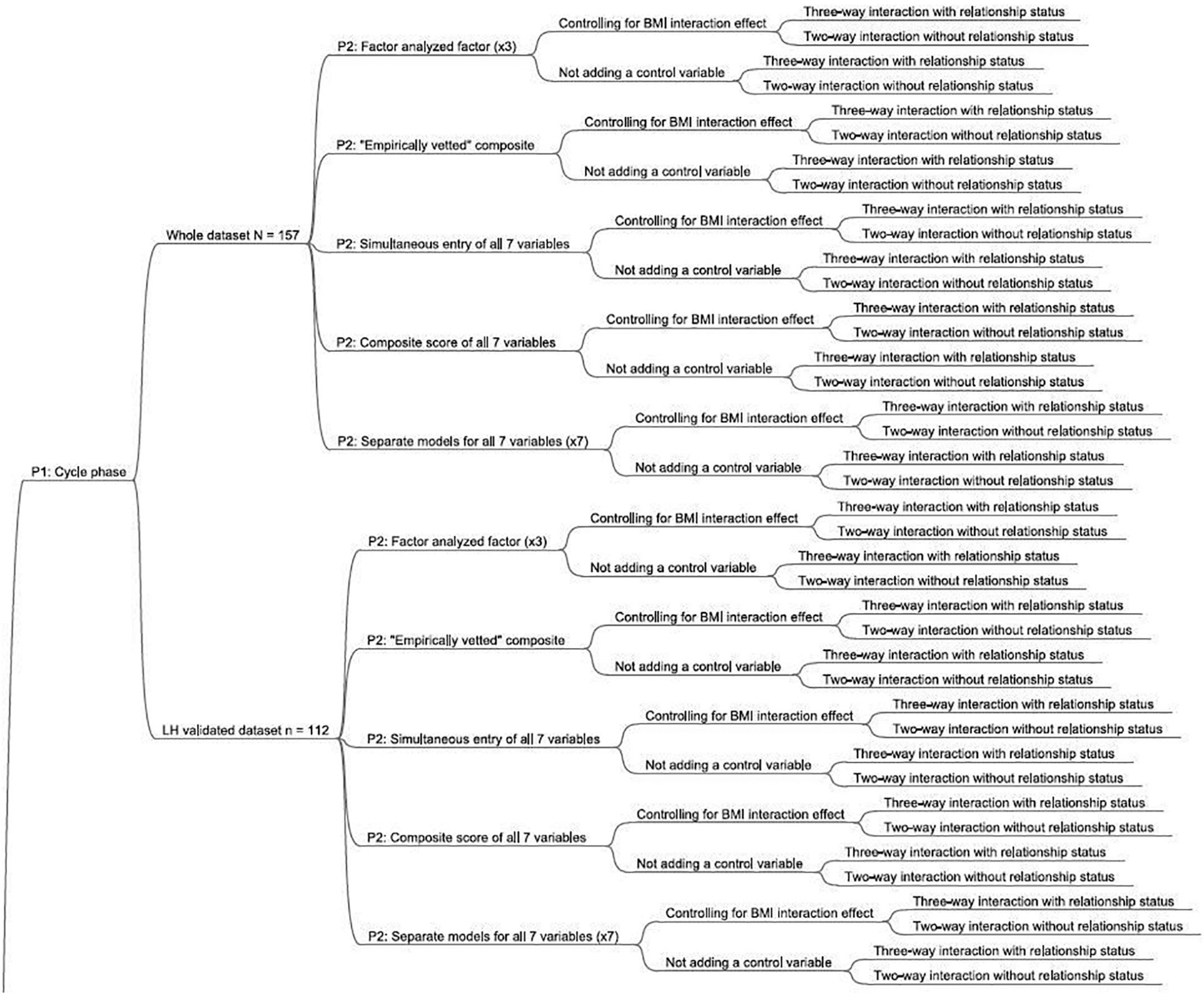
P1: Cycle phase
Whole dataset N = 157
P2: Factor analyzed factor (x3)
Controlling for BMI interaction effect
Three-way interaction with relationship status
Two-way interaction without relationship status
Not adding a control variable
Three-way interaction with relationship status
Two-way interaction without relationship status
P2: "Empirically vetted" composite
Controlling for BMI interaction effect
Three-way interaction with relationship status
Two-way interaction without relationship status
Not adding a control variable
Three-way interaction with relationship status
Two-way interaction without relationship status
P2: Simultaneous entry of all 7 variables
Controlling for BMI interaction effect
Three-way interaction with relationship status
Two-way interaction without relationship status
Not adding a control variable
Three-way interaction with relationship status
Two-way interaction without relationship status
P2: Composite score of all 7 variables
Controlling for BMI interaction effect
Three-way interaction with relationship status
Two-way interaction without relationship status
Not adding a control variable
Three-way interaction with relationship status
Two-way interaction without relationship status
P2: Separate models for all 7 variables (x7)
Controlling for BMI interaction effect
Three-way interaction with relationship status
Two-way interaction without relationship status
Not adding a control variable
Three-way interaction with relationship status
Two-way interaction without relationship status
LH validated dataset n = 112
P2: Factor analyzed factor (x3)
Controlling for BMI interaction effect
Three-way interaction with relationship status
Two-way interaction without relationship status
Not adding a control variable
Three-way interaction with relationship status
Two-way interaction without relationship status
P2: "Empirically vetted" composite
Controlling for BMI interaction effect
Three-way interaction with relationship status
Two-way interaction without relationship status
Not adding a control variable
Three-way interaction with relationship status
Two-way interaction without relationship status
P2: Simultaneous entry of all 7 variables
Controlling for BMI interaction effect
Three-way interaction with relationship status
Two-way interaction without relationship status
Not adding a control variable
Three-way interaction with relationship status
Two-way interaction without relationship status
P2: Composite score of all 7 variables
Controlling for BMI interaction effect
Three-way interaction with relationship status
Two-way interaction without relationship status
Not adding a control variable
Three-way interaction with relationship status
Two-way interaction without relationship status
P2: Separate models for all 7 variables (x7)
Controlling for BMI interaction effect
Three-way interaction with relationship status
Two-way interaction without relationship status
Not adding a control variable
Three-way interaction with relationship status
Two-way interaction without relationship status

Fig. 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 (Fig. S1).

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 1/4th 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 (Fig. 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, Kordsmeyer, 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 et 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 Fig. S1, already led to 416 different models and 1254 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 Fig. 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 1254 resulting p-values, 31 were significant (< 0.05), thus 2.47%. 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 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.

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, Kressel, Joshi, & Louie, 2014). As stated in Gildersleeve et al. (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 et al., 2007; Pawlowski & Jasienska, 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, Galbarczyk, & Jasienska, 2018), bodies (Marcinkowska, Galbarczyk, & Jasienska, 2018; van Stein, Strauß, & Brenk-Franz, 2019), voices (Jünger, Motta-Mena, et al., 2018) 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

[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.

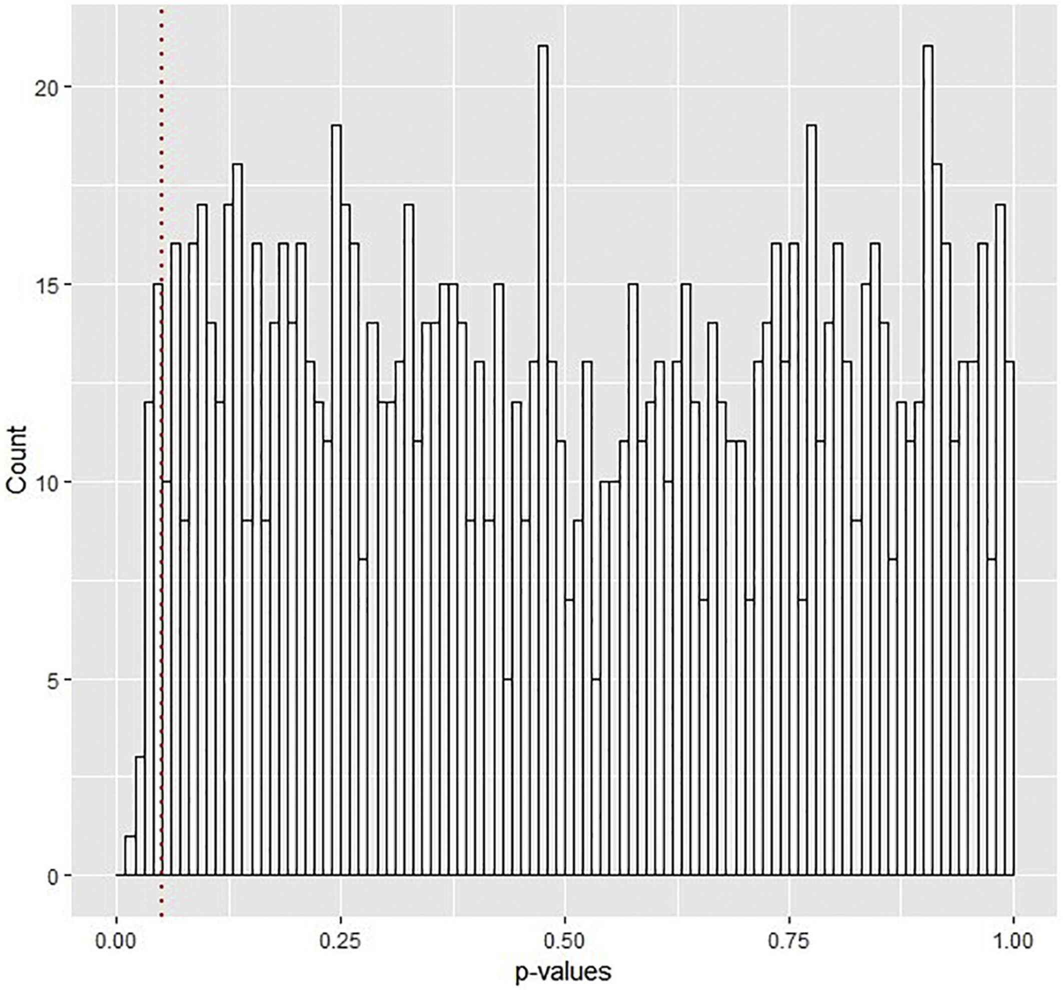

Fig. 3. Histogram displaying the frequency of the 1254 p-values of interest resulting from the multiverse analysis. Note that the red dotted line is at p = .05 and thus separates nominally significant results on the left from nominally non-significant results on the right.

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, Kordsmeyer, et al., 2018; Jünger, Motta-Mena, et al., 2018; 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/

Declaration of Competing Interest

Authors have no competing interest to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.evolhumbehav.2019.08.005.

References

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

Arslan, R. C., Schilling, K. M., Gerlach, T. M., & Penke, L. (2019). Using 26 thousand diary entries to show ovulatory changes in sexual desire and behavior. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000208 in press).

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.yhbeh.2014.04.002.

Gangestad, S. W., Dinh, T., Grebe, N. M., Del Giudice, M., & Emery Thompson, M. (2019). Psychological cycle shifts redux: Revisiting a preregistered study examining preferences for muscularity. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2019.05.005 this issue.

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., 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. https://doi.org/10.1016/j.evolhumbehav.2015.09.001.

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. https://doi.org/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. https://doi.org/10.1037/a0035438.

Higham, J. P. (2016). Field endocrinology of nonhuman primates: Past, present, and future. *Hormones and Behavior, 84*, 145–155. https://doi.org/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. https://doi.org/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. https://doi.org/10.1177/0956797618760197.

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*, 412–423. https://doi.org/10.1016/j.evolhumbehav.2018.03.007.

Jünger, J., Motta-Mena, N. V., Cardenas, R., Bailey, D., Rosenfield, K. A., Schild, C., ... Puts, D. A. (2018). Do women's preferences for masculine voices shift across the ovulatory cycle? *Hormones and Behavior, 106*, 122–134. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.yhbeh.2007.03.006.

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 (doi: 10.106/j.psyneuen.2017.11.007).

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

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. https://doi.org/10.1037/a0028380.

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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1177/1745691616658637.

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. https://doi.org/10.1177/1474704919848116.

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 fromhttps://psyarxiv.com/7g3xchttps://doi.org/10.17605/OSF.IO/7G3XC.

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. https://doi.org/10.1177/1754073914523073.

Evolution and Human Behavior 40 (2019) 526–530

Contents lists available at ScienceDirect

Evolution and Human Behavior

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

Commentary

On the use of log transformations when testing hormonal predictors of cycle phase shifts: Commentary on Gangestad, Dinh, Grebe, Del Giudice, and Emery Thompson (2019)

James R. Roney

Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106, United States of America.

ARTICLE INFO

Keywords:
Menstrual cycle
Estradiol
Progesterone
Log transformations

1. Introduction

Gangestad, Dinh, Grebe, Del Giudice, and Emery Thompson (this issue) argue that Jünger, Kordsmeyer, Gerlach, and Penke (2018) were "premature" to conclude that their own data failed to support an ovulatory shift whereby women experience greater attraction to men with more masculine bodies on days when conception is possible relative to other cycle days. In a re-analysis of Jünger et al.'s open data, they argue instead that one specific pattern supports such a cycle shift: a three-way interaction between within-women shifts in women's log transformed estradiol-to-progesterone (EP) ratio, their relationship status, and their preferences for male stimuli high in strength/muscularity. Stern, Arslan, Gerlach, and Penke (2019), in a response to the Gangestad et al. re-analysis, argue via a multiverse analysis that this effect is not robust to alternative yet reasonable data analytic decisions, including decisions regarding the log transformation of hormone ratios (that is, the reported three-way interaction is not significant without the log transformation of the EP ratio). Here, to avoid redundancy with the other commentaries, I will focus mainly on this issue of log transformation, which may have broader implications for data analyses in behavioral endocrinology, but is also directly relevant to the robustness of the Gangestad et al. findings. I will then conclude with a brief discussion of the theoretical implications of the positive result proposed by Gangestad et al., and suggest that greater clarity regarding the theories that Gangestad et al. are testing is necessary to ensure that their positions are falsifiable.

2. Mechanisms, indexes, and logs

Jünger et al. (2018) in their original article were testing hypotheses about cycle phase effects: in particular, the "good genes ovulatory shift hypothesis" that predicts stronger attraction to putative good genes indicators during the fertile window of the menstrual cycle (i.e. the cycle days when conception is possible) than on other cycle days. They report null results for fertile window shifts in preferences for masculine bodies, even for a subsample of participants who had positive luteinizing hormone (LH) tests the timing of which should have placed most of the women within the putative fertile window. Gangestad et al., however, shifted the focus of their analyses away from cycle phase to effects of ovarian hormones that are the presumed mechanisms that implement fertile window shifts in psychology and behavior. They justified this by arguing that positive LH tests alone may not distinguish between cycles that are ovulatory but sub-fertile vs. ovulatory with high fertility (whereas hormones should differ across such cycles), and by arguing that the power to detect cycle phase shifts should be greater when directly assessing the mechanisms that implement such shifts. Setting aside the merits of these arguments (for a counterargument based on measurement error associated with salivary hormones, see Stern et al., 2019), the important point here is that the underlying theoretical issues concern fertile window shifts in preferences, and thus any hormone variables that are used in data analyses should be good indexes of fertile window timing.

E-mail address: roney@psych.ucsb.edu.

https://doi.org/10.1016/j.evolhumbehav.2019.08.006
Received 22 August 2019; Accepted 23 August 2019

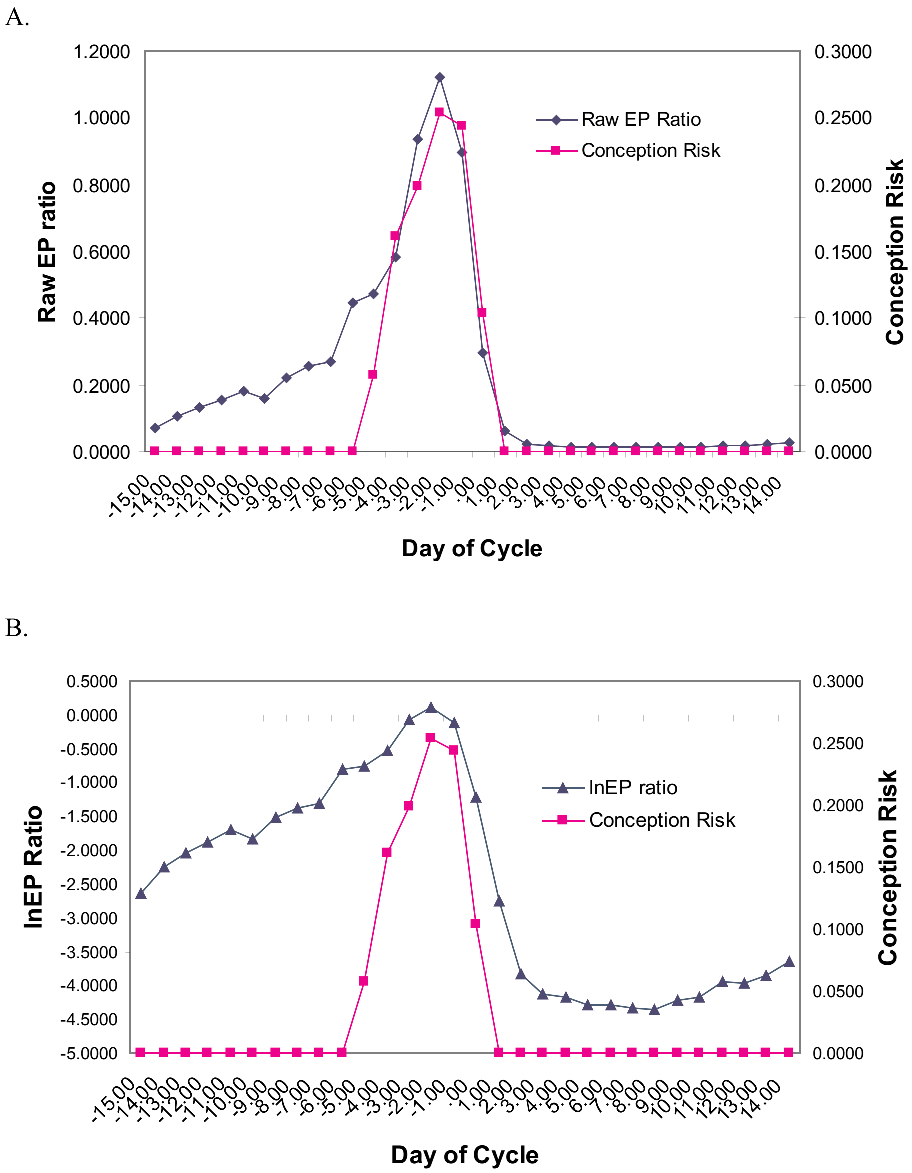

Fig. 1. Raw (A) and log transformed (B) EP ratios plotted with conception risk estimates relative to days of the menstrual cycle. Day zero represents the estimated day of ovulation, with days before ovulation numbered negatively and after ovulation numbered positively. EP = estradiol to progesterone.

EP ratio is known to be such an index, given the secretion patterns of the two hormones in ovulatory cycles (see below). Gangestad et al. argued forcefully that the natural log transformed EP ratio (hereafter lnEP) is preferable to the raw EP ratio for use in hypothesis tests of cycle phase shifts, and in fact reported results using only lnEP and not raw EP ratio. In what follows, though, I will suggest that raw EP ratio is actually the better index of conception risk and fertile window timing, which at least raises questions regarding the appropriateness of the Gangestad et al. re-analyses of the Jünger et al. (2018) data. I will then answer some of Gangestad et al.'s counterarguments to a version of this argument that I communicated to them in my signed review of their article.

Studies with precise estimates of ovulatory timing have established that conception in humans is only possible in a restricted window of days from at most five days before the day of ovulation through the day of ovulation itself (e.g., Wilcox, Weinberg, & Baird, 1998). Within this fertile window, however, risk of conception also varies considerably; Shirazi, Jones, Roney, DeBruine, and Puts (2019) recently computed weighted estimates of conception probability on each fertile window day using data from prior studies. I used these estimates to plot conception risk against day of the menstrual cycle, with days outside of the human fertile window assigned conception risk of zero (see Fig. 1). I also used data from Stricker et al. (2006) on mean serum concentrations of estradiol and progesterone on individual cycle days to compute daily EP ratios (E/P after conversion to pmol/L) and then lnEP ratios. Raw and log transformed EP ratios are plotted on the same graphs with conception risk in Fig. 1a and b, respectively.

It can be seen from these figures that raw EP ratio is a remarkably good index of conception risk in the human menstrual cycle. lnEP ratio appears to be a much poorer index of conception probability. Confirming these visual intuitions, conception risk correlates more strongly with raw EP ratio, $r = 0.92$, $R^2 = 0.85$ than it does with lnEP ratio, $r = 0.68$, $R^2 = 0.46$. Notice that raw EP ratio should track between-cycle differences in fertility that Gangestad et al. cited as a reason for focusing on hormones over cycle phase itself (that is, the peaks in this ratio will be higher in fertile windows with higher estradiol), but also has the advantage of changing very little during the luteal phase when conception risk is uniformly zero, which is less true of lnEP ratio. Thus, at first glance, raw EP ratio is the better proxy for whatever hormonal mechanisms may regulate potential fertile window shifts in outcome measures.

Gangestad et al. cite two broad reasons that lnEP ratio is a preferable variable even if raw EP ratio correlates more highly with conception risk. First, highlighting the fact that lnEP = ln(E) – ln(P), they point out that "…ln(E/P) captures equal but opposite joint additive contributions of ln(E) and ln(P). (It constrains the regression weights of ln(E) and ln(P) to be identical in magnitude but opposite in sign.)" Gangestad et al. (p. 4). Raw EP ratio, on the other hand, does not have this straightforward interpretation, and in addition to additive effects of the hormones also "reflects complex non-linear main effects and interactions" (fn 7, Gangestad et al., p. 4). Furthermore, EP ratio is not symmetric to PE ratio. Second, they suggest that hormone effects are often non-linear since receptor saturation implies that increasing hormone concentrations will have diminishing effects. The first argument seems especially important here, since Gangestad et al. are essentially suggesting that equal but opposite additive effects of estradiol and progesterone is a reasonable model of cycle phase dynamics associated with potential fertile window shifts in outcome measures. But is this in fact a good model?

Let's consider evidence regarding the neural mechanisms related to effects of these two hormones in nonhuman species. Such effects are clearly dynamic and sequential, and not easily captured by simple additive influences. Estradiol administration causes induction of both estradiol and progesterone receptors over time (Carter, 1992; Parsons, Maclusky, Krey, Pfaff, & McEwen, 1980; Sá & Fonseca, 2017; Siegel, Senatore, Rogers, & Ahdieh, 1989), thus adjusting brain responsiveness to changing production of these hormones. Estradiol also promotes the formation and enhanced thickness of dendritic spines associated with synapse formation in regions such as the ventromedial hypothalamus (VMH) and the hippocampus, whereas progesterone has been shown to promote the loss of these spines, with such effects also observed in natural cycles and temporally correlated with changes in sexual receptivity (e.g., McEwen & Woolley, 1994; Woolley & McEwen, 1993). Through such genomic effects on receptor expression and synapse formation, as well as more immediate membrane-mediated effects, rising estradiol during the follicular phase may have time-lagged, cumulative, and also short-term effects on sexual receptivity, whereas rising progesterone after ovulation may reverse some of the estradiol-mediated synaptic changes and thereby inhibit sexual motivation (see Flanagan-Cato, 2000; Kow & Pfaff, 2004; McEwen & Woolley, 1994).

Given these dynamic and sequential effects of estradiol and progesterone, are equal but opposite effects of the two hormones likely to accurately model their influences? A unit change in estradiol may in fact have a much different effect during the late follicular phase given one configuration of synaptic connections within hypothalamic networks than it does during the luteal phase when progesterone has altered those connections. Likewise, progesterone may have virtually no effects at all in neurons within which estradiol has not yet induced progesterone receptors. A simple model in which neurons read out immediate estradiol and progesterone concentrations and then produce outputs proportional to a difference between those concentrations thus appears unrealistic given what we know about the dynamic influences of these signals in natural estrous and menstrual cycles.[1] Gangestad et al. criticized the use of raw EP ratio since much of its variance (at least in the Jünger et al. dataset) reflects "complex non-linear main effects [of] and interactions" between the two hormones, but it is unclear that this is problematic when the neural effects of the hormones also reflect complex temporal patterns whereby estradiol and progesterone interact to influence synaptic connectivity and neuron firing thresholds. Thus, although it is a nice mathematical property of lnEP ratio that it "is explained by simple additive effects of ln(E) and ln(P)" (Gangestad et al. fn 7, p. 4), this alone does not recommend its use over raw EP ratio in tests of fertile window shifts in preferences if such additive effects do not accurately model neural effects of the hormones and if raw EP ratio is simply a better proxy for changes in conception risk across full menstrual cycles.

Importantly, I am not arguing that brain mechanisms read out raw EP ratio at any given moment and then respond proportionally. Rather, raw EP ratio is a good *index* of the outcomes of complex, temporal sequences of hormonal influences, and thus of brain states associated with high vs. low conception risk.[2] In Fig. 1a, a high EP ratio in the late follicular phase indexes brain states associated with high conception risk in part because of cumulative effects of estrogen priming on prior days, whereas a uniformly low EP ratio in the luteal phase indexes brain states associated with low conception risk, potentially because progesterone has reversed some of the synaptic effects caused by follicular phase estrogen priming. Thus, although the actual mechanisms for cycle phase shifts involve sequential and interactive influences of estradiol, progesterone, and perhaps other signals, raw EP ratio appears to provide a good index of within-cycle shifts in conception probability.

There are two implications of this discussion of log transformations that are relevant here. First, the fact that the EP ratio x relationship status x strength/muscularity interaction is not significant when substituting raw EP ratio for lnEP ratio (Stern et al., 2019) raises important doubts about whether this interaction is robust. Given that raw EP ratio provides the better index of conception risk (Fig. 1), a null result using this ratio argues against an ovulatory shift in preferences for body muscularity. EP and lnEP ratio are positively correlated, of course, as Gangestad et al. note, and both are positively correlated with estimates of conception risk, but it is troubling that the reported interaction is significant with only one measure and not the other.

[1] Gangestad et al. imply that some of my prior publications (Roney & Simmons, 2013, 2017) illustrate that additive but opposite effects of estradiol and progesterone fully explain fertile window shifts in outcome variables; for sexual motivation, for instance, they wrote: "Roney & Simmons, 2013, found that, with ovarian hormone levels controlled, there was no significant residual effect of estimated conception risk" (p. 1). This is not completely accurate for this study. Cognizant of the sequential effects of estradiol and progesterone, we reported follow-up analyses in this paper in which we considered the follicular phase and the fertile window plus luteal phase separately. For the follicular phase, we found only a slight reduction in effect size for the effect of fertile window timing on sexual desire when adding hormone variables to the same model, and thus we were unable to explain much of the mid-cycle rise in desire via the hormone variables. We speculated that other signals may combine with estradiol to explain this rise (e.g., LH or oxytocin), though more complex cumulative effects of estradiol that were not captured in our regression models may also be relevant. The fall in desire from the fertile window to the luteal phase, on the other hand, did appear to be statistically explained by the rise in progesterone.

[2] Note also that progesterone to estradiol (PE) ratio is not a comparably good index of conception risk, largely because it is relatively invariant when conception risk changes rapidly during the fertile window, but then has greater variability during the luteal phase when conception risk is constant at zero. Using the Stricker et al. (2006) data, PE ratio correlates with conception risk as depicted in Fig. 1 at $r = -0.44$, $R^2 = 0.19$. Thus, while it is true the EP ratio does not co-vary perfectly negatively with PE ratio, as pointed out by Gangestad et al. (see section 27 of SOM), it is empirically the case that the former is a good index of conception risk and the latter is not.

Second, the issues discussed above suggest caution regarding the general adoption of lnEP ratio as a hormonal measure of cycle phase shifts in human behavioral endocrinology. As a proxy for conception risk, this measure appears to perform relatively poorly. More generally, it is not entirely clear that log transformations of individual hormones are always advisable when testing the relationships between within-women changes in hormones and other outcome variables. Receptor saturation arguments may not be generally applicable when measuring ovarian hormones in natural cycles, since, as described above, rising estrogen can increase hormone receptor expression. Furthermore, estradiol administration within the physiological range produced a linear dose-response effect on hippocampal brain activation in young women (Bayer, Gläscher, Finsterbusch, Schulte, & Sommer, 2018), demonstrating that response functions to changing hormones cannot be assumed to be logarithmic in all cases. Second, log transformations when applied to change scores (which are relevant to within-women data analyses) produce effects analogous to percent changes, such that small changes from low baselines become equivalent to large changes from higher baselines. Yet, it is not at all clear that brain mechanisms should be designed to respond to hormone variability in this way. Higher fecundity menstrual cycles exhibit higher estradiol production across most of the menstrual cycle relative to lower fecundity cycles from the same women (e.g., Lipson & Ellison, 1996). This means that large ovulatory increases in estradiol from higher follicular phase baselines are associated with greater fecundity than are small increases from lower baselines; logarithmic transformations of estradiol, however, can eliminate differences in absolute change scores in such cases despite the fact that these differences are biologically meaningful.[3] Based on these types of considerations, much more evidence appears necessary before adopting log transformations as standard practice for the analysis of hormone data within the context of human menstrual cycle research.

3. Effects, theories, and falsification

Gangestad et al. conclude that Junger et al.'s (2018) "null conclusions" are "premature" (p. 1), but an important issue concerns the question of how empirical conclusions are related to specific theories. Jünger et al. (2018) set out to test predictions of the good genes ovulatory shift hypothesis, and in that light, overall null conclusions may be justified from their data. As Stern et al. describe more fully in their commentary, even if one concedes the presence of the three-way interaction proposed by Gangestad et al., the overall pattern of results does not appear completely consistent with the good genes ovulatory shift hypothesis. The two-way interaction between strength/muscularity and lnEP ratio was not significant, and even the three-way interaction involving relationship status was largely driven by a negative relationship between changes in lnEP ratio and attractiveness ratings of more muscular men among single women, which is an effect that if anything is opposite to predictions from the ovulatory shift hypothesis. In addition, one could postulate from prior findings in the literature (e.g., Penton-Voak et al., 1999) that preference shifts should be found only for ratings of short-term or sexual attractiveness, but the three-way interaction reported by Gangestad et al. was also found for ratings of long-term attractiveness.

Gangestad et al. discuss the interpretation of their proposed three-way interaction effect in Section 5.5, and are agnostic regarding what it means. They write that it could "potentially be consistent with a good genes framework," but that other explanations should be considered, "including Type I error, conjectures that non-conceptive sex plays special roles in partnered women (Grebe, Gangestad, Garver-Apgar, & Thornhill, 2013), and other perspectives on human mating (Emery Thompson & Muller, 2016)" (p. 13). Yet, it is not clear what predictions each of the latter two positions make regarding hormonal predictors of preferences for body masculinity, and thus exactly what findings constitute effects that support or refute specific theoretical positions. This in turn raises questions regarding what we should consider an "effect" at all. There is a very large set of potential statistical tests that might demonstrate a significant effect for attractiveness ratings involving interactions with cycle phase or hormonal variables, but the tests we should be focused on depend on how specific effects relate to specific theoretical positions. If "ovulatory shift" theories become so vaguely specified that any significant complex interaction involving a cycle phase or hormonal variable can be considered confirmatory evidence, then the theories in question seem at risk of being unfalsifiable. Jünger et al. (2018) set out to test a group of hypotheses derived from the good genes ovulatory shift hypothesis, and their characterization of the overall pattern of results as inconsistent with this hypothesis seems reasonable, even if the specific three-way interaction reported in the Gangestad et al. re-analysis proves to be robust.

Acknowledgements

I thank James Peter Roney for helpful comments on a version of this commentary.

References

Bayer, J., Gläscher, J., Finsterbusch, J., Schulte, L. H., & Sommer, T. (2018). Linear and inverted U-shaped dose-response functions describe estrogen effects on hippocampal activity in young women. *Nature Communications, 9*(1), 1–12. https://doi.org/10.1038/s41467-018-03679-x.

Carter, C. S. (1992). Neuroendocrinology of sexual behavior in the female. In J. B. Becker, S. M. Breedlove, & D. Crews (Eds.). *Behavioral endocrinology* (pp. 72–95). Cambridge, MA: MIT Press.

Flanagan-Cato, L. M. (2000). Estrogen-induced Remodeling of hypothalamic neural circuitry. *Frontiers in Neuroendocrinology, 21*(4), 309–329. https://doi.org/10.1006/frne.2000.0204.

Gangestad, S. W., Dinh, T., Grebe, N. M., Del Giudice, M., & Emery Thompson, M. (2019). Psychological cycle shifts redux: Revisiting a preregistered study examining preferences for muscularity. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2019.05.005 this issue.

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. https://doi.org/10.1016/j.evolhumbehav.2018.03.007.

Kow, L.-M., & Pfaff, D. W. (2004). The membrane actions of estrogens can potentiate their lordosis behavior-facilitating genomic actions. *Proceedings of the National Academy of Sciences, 101*(33), 12354–12357. https://doi.org/10.1073/pnas.0404889101.

Lipson, S. F., & Ellison, P. T. (1996). Comparison of salivary steroid profiles in naturally occurring conception and non-conception cycles. *Human Reproduction (Oxford, England), 11*(10), 2090–2096.

McEwen, B. S., & Woolley, C. S. (1994). Estradiol and progesterone regulate neuronal structure and synaptic connectivity in adult as well as developing brain. *Experimental Gerontology, 29*(3), 431–436. https://doi.org/10.1016/0531-5565(94)90022-1.

Parsons, B., Maclusky, N. J., Krey, L., Pfaff, D. W., & McEwen, B. S. (1980). The temporal relationship between Estrogen-inducible progestin receptors in the female rat brain and the time course of Estrogen activation of mating behavior. *Endocrinology, 107*(3), 774–779. https://doi.org/10.1210/endo-107-3-774.

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), 741–742. https://doi.org/10.1038/21557.

Roney, J. R., & Simmons, Z. L. (2013). Hormonal predictors of sexual motivation in natural menstrual cycles. *Hormones and Behavior, 63*(4), 636–645. https://doi.org/10.1016/j.yhbeh.2013.02.013.

Roney, J. R., & Simmons, Z. L. (2017). Ovarian hormone fluctuations predict within-cycle shifts in women's food intake. *Hormones and Behavior, 90*, 8–14. https://doi.org/10.1016/j.yhbeh.2017.01.009.

Sá, S. I., & Fonseca, B. M. (2017). Dynamics of progesterone and estrogen receptor alpha in the ventromedial hypothalamus. *The Journal of Endocrinology, 233*(2), 197–207. https://doi.org/10.1530/JOE-16-0663.

Shirazi, T. N., Jones, B. C., Roney, J. R., DeBruine, L. M., & Puts, D. A. (2019). Conception risk affects in-pair and extrapair desire similarly: A comment on Shimoda et al. (2018). *Behavioral Ecology, 30*(4), e6–e7. https://doi.org/10.1093/beheco/arz056.

Siegel, H. I., Senatore, A., Rogers, S., & Ahdieh, H. B. (1989). Sexual receptivity in

[3] A concrete example may help to illustrate this point. Assume two women, A and B, have differing estradiol concentrations in measured menstrual cycles, with measurements on days -8 and -1 relative to the day of ovulation. I used mean estradiol concentrations (pmol/L) on these cycle days from Stricker et al. (2006) and divided them in half for woman A and doubled them for woman B. Thus, the values are: 85.17 and 457.42 on the respective cycle days for A, and 340.68 and 1829.68 for B. Raw hormone difference scores across the two cycle days are 372.25 and 1489 for A and B, respectively. With log transformations of all hormone values, however, the difference scores are now an identical 0.73 for both A and B. Data analyses on log transformed data will thus treat A and B as having identical changes in hormones.

hamsters: Brain nuclear estrogen and cytosolic progestin receptors after single and multiple steroid treatments and during the estrous cycle. *Hormones and Behavior, 23*(2), 173–184. https://doi.org/10.1016/0018-506X(89)90058-5.

Stern, J., Arslan, R. C., Gerlach, T. M., & Penke, L. (2019). Using multiverse analysis to show that cycle shifts in preferences for men's bodies are not robust. *Evolution and Human Behavior* this issue.

Stricker, R., Eberhart, R., Chevailler, M.-C., Quinn, F. A., Bischof, P., & Stricker, R. (2006). Establishment of detailed reference values for luteinizing hormone, follicle stimulating hormone, estradiol, and progesterone during different phases of the menstrual cycle on the Abbott ARCHITECT analyzer. *Clinical Chemistry and Laboratory Medicine, 44*(7), 883–887. https://doi.org/10.1515/CCLM.2006.160.

Wilcox, A. J., Weinberg, C. R., & Baird, D. D. (1998). Post-ovulatory ageing of the human oocyte and embryo failure. *Human Reproduction (Oxford, England), 13*(2), 394–397.

Woolley, C. S., & McEwen, B. S. (1993). Roles of estradiol and progesterone in regulation of hippocampal dendritic spine density during the estrous cycle in the rat. *The Journal of Comparative Neurology, 336*(2), 293–306. https://doi.org/10.1002/cne.903360210.

Evolution and Human Behavior 40 (2019) 531–532

Contents lists available at ScienceDirect

Evolution and Human Behavior

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

Commentary

Assessing the evidentiary value of secondary data analyses: A commentary on Gangestad, Dinh, Grebe, Del Giudice, and Thompson (2019)

Benedict Christopher Jones[a,*], Urszula Maria Marcinkowska[b], Lisa Marie DeBruine[a]

[a] *Institute of Neuroscience & Psychology, University of Glasgow, USA*
[b] *Yale Reproductive Ecology Laboratory, Yale University, Jagiellonian University Medical College, USA*

Secondary data analyses (analyses of open data from published studies) can play a critical role in hypothesis generation and in maximizing the contribution of collected data to the accumulation of scientific knowledge. However, assessing the evidentiary value of results from secondary data analyses is often challenging because analytical decisions can be biased by knowledge of the results of (and analytical choices made in) the original study and by unacknowledged exploratory analyses of open data sets (Scott & Kline, 2019; Weston, Ritchie, Rohrer, & Przybylski, 2018). Using the secondary data analyses reported by Gangestad et al. (2019) as a case study, we outline several approaches that, if implemented, would allow readers to assess the evidentiary value of results from secondary data analyses with greater confidence.

Jünger, Kordsmeyer, Gerlach, and Penke (2018) reported results of a longitudinal study testing for evidence that women's preferences for masculine men's body shapes track changes in fertility. They found no evidence for significant effects of fertility on preferences for markers of masculinity, such as muscularity and height, adding to a growing literature that finds little evidence that within-woman changes in preferences for putative male cues of good genes reliably track within-woman changes in conception risk or steroid hormone levels (Dixson et al., 2018; Jones et al., 2018; Jünger et al., 2018; Marcinkowska et al., 2016, Marcinkowska, Galbarczyk, & Jasienska, 2018, Marcinkowska, Kaminski, Little, & Jasienska, 2018, Marcinkowska, Helle, Jones, & Jasienska, 2019; Stern, Gerlach, & Penke, 2018).

Gangestad et al. (2019) recently published alternative analyses of Jünger, Kordsmeyer, et al. (2018) open data. They argued that the results of these reanalyses suggest that women's preferences for muscularity do indeed track changes in sex hormones (specifically, progesterone). Gangestad et al. highlighted that their secondary analyses were closely modelled on their own preregistered analysis plans for an ongoing study by Gangestad and colleagues. This ongoing study had a similar design to the study reported by Jünger, Kordsmeyer, et al. (2018). However, since Jünger, Kordsmeyer, et al., 2018 paper was cited in Gangestad et al.'s preregistration (https://osf.io/zbktu/), that preregistration was not blind to the results and analytical choices reported in Jünger, Kordsmeyer, et al. (2018) paper. In addition, Gangestad et al.'s preregistration cites Jünger, Kordsmeyer, et al. (2018) for the same effect of progesterone on body shape preferences that Gangestad et al. (2019) report in their reanalyses, but that was not evident in Jünger et al.'s original analyses.

Because of the above, using Gangestad et al.'s preregistration as the basis for their reanalyses of Jünger, Kordsmeyer, et al. (2018) data would not have sufficiently guarded against biases that may have been introduced by knowledge of Jünger, Kordsmeyer, et al. (2018) analytical choices and results. In other words, the claim that Gangestad et al.'s results have high evidentiary value because their analyses were closely modelled on a preregistered analysis plan, is circular. Therefore, locating Gangestad et al.'s reanalyses on the continuum between confirmatory (where statistical significance can be interpreted as indicating that a given result has high evidentiary value) and exploratory (where statistical significance cannot be interpreted as indicating that a given result has high evidentiary value) is not straightforward.

As is the case in many other areas, open data and analysis code are becoming the standard in research on menstrual cycle and mate preferences. How can the field ensure that readers can assess the evidentiary value of secondary data analyses with confidence? Below, we highlight four possible solutions to this problem. The solutions we outline are not intended as an exhaustive list and are also not necessarily mutually exclusive.

The first solution is to use open data for hypothesis generation. In such cases, the results should be clearly labelled as exploratory if published (see Weston et al., 2018 for a discussion of the importance of this type of transparency when reporting secondary data analyses). Alternatively, authors could wait until confirmatory analyses have been carried out on a newly collected data set before publication. That the preregistration Gangestad et al. modelled their reanalyses on includes a direct reference to the results of their reanalyses of Jünger, Kordsmeyer, et al. (2018) data means it is misleading to imply that preregistering their analysis plan increases the evidentiary value of Gangestad et al.'s secondary data analyses. However, it is encouraging that Gangestad and colleagues plan to replicate their reanalysis of Jünger et al.'s data on a new data set that they are currently collecting. Importantly, the open analysis code they have included with their reanalysis means it

[*] Corresponding author.
E-mail address: ben.jones@glasgow.ac.uk (B.C. Jones).

https://doi.org/10.1016/j.evolhumbehav.2019.08.004
Received 5 August 2019; Received in revised form 14 August 2019; Accepted 15 August 2019

will be relatively straightforward for readers of this forthcoming work to directly compare those with the ones reported by Gangestad et al. (2019).

The second solution is to use open data for direct replication of new results (i.e., to confirm results of other studies by replicating their analyses on other open data). A recent example of this approach comes from DeBruine, Hahn, and Jones (2019), who used open data from Jones et al. (2018) to directly replicate Marcinkowska, Kaminski, et al. (2018) analyses of (and results for) combined effects of average progesterone levels and partnership status on women's preferences for male facial masculinity.

The third solution is to use specification curve analyses (Simonsohn, Simmons, & Nelson, 2015) or the closely related technique multiverse analyses (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). These methods indicate how robust results are across a range of reasonable analytical plans. For example, this approach has recently proven effective in analyses of the possible effects of social media use on life satisfaction (Orben, Dienlin, & Przybylski, 2019), demonstrating that such effects are typically small and very sensitive to the specific analytical choices made. Indeed, Stern, Arslan, Gerlach, and Penke's (2019) analyses demonstrate that the significance of the effects that Gangestad et al. reported is extremely sensitive to specific analytical decisions made, suggesting they are not robust and have low evidentiary value.

A fourth potential solution is for the field to create a data management infrastructure that would allow large data sets to be made open in phases (Scott & Kline, 2019). With this solution, some of the data from a study would be made open for exploratory analyses immediately. The remainder (i.e., an independent or 'hold out' sample) would then be made open at a later date for preregistered confirmatory analyses based on the earlier exploratory analyses. This approach has been employed successfully in other research areas (e.g., physics) and is being adopted by some large-scale multisite research initiatives (e.g., recent developments with the Psychological Science Accelerator or the Attitudes, Identities, and Individual Differences Study). Indeed, combining this approach with specification curve analyses may be optimal for guarding against biases.

In closing, we reiterate that secondary data analyses are essential for hypothesis generation and maximizing the contribution that published data can make to the accumulation of scientific knowledge. However, to make a substantial contribution requires that people are able to assess the evidentiary value of secondary data analyses, both accurately and confidently. We encourage researchers, readers, and editors to carefully consider how secondary data analyses reported in papers and presented at conferences were conducted, described, and interpreted when assessing the evidentiary value of their results. We believe that considering the solutions described above would allow the evidentiary value of further secondary analyses of data to be assessed more confidently.

References

DeBruine, L. M., Hahn, A. C., & Jones, B. C. (2019). Does the interaction between partnership status and average progesterone level reliably predict women's preferences for facial masculinity? *Hormones and Behavior, 107*, 80–82. https://doi.org/10.1016/j.yhbeh.2018.12.004.

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.

Gangestad, S. W., Dinh, T., Grebe, N. M., Del Giudice, M., & Thompson, M. E. (2019). Psychological cycle shifts redux: Revisiting a preregistered study examining preferences for muscularity. *Evolution and Human Behavior* (this issue).

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.

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.

Jünger, J., Motta-Mena, N. V., Cardenas, R., Bailey, D., Rosenfield, K. A., Schild, C., ... Puts, D. A. (2018). Do women's preferences for masculine voices shift across the ovulatory cycle? *Hormones and Behavior, 106*, 122–134.

Marcinkowska, U. M., Ellison, P. T., Galbarczyk, A., Milkowska, K., Pawlowski, B., Thune, I., & Jasienska, G. (2016). Lack of support for relation between woman's masculinity preference, estradiol level and mating context. *Hormones and Behavior, 78*(1–7).

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.

Marcinkowska, U. M., Helle, S., Jones, B. C., & Jasienska, G. (2019). Does testosterone predict women's preference for facial masculinity? *PLoS ONE, 14*(2), e0210636.

Marcinkowska, U. M., Kaminski, G., Little, A. C., & Jasienska, G. (2018). Average ovarian hormone levels, rather than daily values and their fluctuations, are related to facial preferences among women. *Hormones and Behavior, 102*, 114–119.

Orben, A., Dienlin, T., & Przybylski, A. K. (2019). Social media's enduring effect on adolescent life satisfaction. *Proceedings of the National Academy of Sciences, 116*(21), 10226–10228.

Scott, K. M., & Kline, M. (2019). Enabling confirmatory secondary data analysis by logging data checkout. *Advances in Methods and Practices in Psychological Science, 2*(1), 45–54.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). *Specification curve: Descriptive and inferential statistics on all reasonable specifications.* Available at SSRN 2694998.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*, 702–712.

Stern, J., Arslan, R. C., Gerlach, T. M., & Penke, L. (2019). *No robust evidence for cycle shifts in preferences for men's bodies: a multiverse analysis. Evolution and Human Behavior.*.

Stern, J., Gerlach, T. M., & Penke, L. (2018). *Probing ovulatory cycle shifts in women's preferences for men's behaviors.* PsyArxiv.

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2018). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*2515245919848684.

Evolution and Human Behavior 40 (2019) 533–536

Contents lists available at ScienceDirect

Evolution and Human Behavior

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

Commentary

A comparative perspective on measures of cycle phase, and how they relate to cues, signals, and mating behavior: A commentary on Gangestad, Dinh, Grebe, Del Giudice, and Emery Thompson (2019)

James P. Higham

Dept. of Anthropology, New York University, 25 Waverly Place, New York, NY 10003, USA

1. Introduction

Studies investigating the relationships between ovulatory cycles, their potential cues or signals, and female receptivity, proceptivity, and mate choice preferences have been foundational within the field of evolutionary psychology. This includes many studies that have investigated whether changes in cues or signals, such as changes in scent (Gangestad & Thornhill, 1998; Thornhill et al., 2003) and voice pitch (Bryant & Haselton, 2009; Fischer et al., 2011), or in behavior such as clothing choice (Haselton, Mortezaie, Pillsworth, Bleske-Rechek, & Frederick, 2007) vary across the menstrual cycle and are potentially informative about the timing of the fertile phase. One particular thread of studies has investigated women's sexual motivation and mate choice preferences across the cycle. According to the sexual strategies theory (Buss & Schmitt, 1993; Gangestad & Simpson, 2000), there are differences in the short-term and long-term mate-choice priorities of men and women. This theory has led to the proposition of the dual mating strategy hypothesis, which postulates that heterosexual women exhibit different mate preferences at different times of their cycle (reviewed in Jones, Hahn, & Debruine, 2019). According to this hypothesis, heterosexual women's preferences in traits of *long-term* partners do not change across the menstrual cycle (e.g., Gangestad, Simpson, Cousins, Garver-Apgar, & Christensen, 2004; Penton-Voak et al., 1999). In contrast, preferences for *short-term* partners are proposed to shift cyclically. Under this hypothesis, when heterosexual women are in the most fertile phase of their cycle around ovulation, they prefer short-term partners who exhibit traits suggested to indicate “good genes” (Gangestad, Garver-Apgar, Simpson, & Cousins, 2007). Consistent with this, studies have shown that during their fertile phase relative to other phases of the cycle, heterosexual women prefer the scent of men who are more socially dominant (Havlíček, Roberts, & Flegr, 2005), men with more masculine faces (Penton-Voak et al., 1999), men with deeper voices (Puts, 2005), and men who display greater social presence and dominance (Gangestad et al., 2004).

With the ‘replication crisis’ (Pashler & Wagenmakers, 2012), in which replications of oft-cited studies have failed to find the originally-reported effects, evolutionary psychology has commendably joined other subfields of psychology in taking on the mantel of improved empirical rigor. This has involved a number of specific efforts by researchers in the field, including pre-registration of studies, and the use of much larger sample sizes. A number of such studies have failed to replicate previously demonstrated effects of the cycle phase on heterosexual women's mate choice preferences. For example, Jones et al. (2018), showed that women's preferences for facial masculinity do not change according to changes in concentrations of estradiol and progesterone across the cycle. Different authors of meta-analyses of earlier studies on menstrual cycles and mate preferences have come to different conclusions. While Gildersleeve, Haselton, and Fales (2014a, 2014b) concluded that such effects are robust, Wood and Carden (2014),Wood, Kressel, Joshi, and Louie (2014) concluded that they were not.

Recently, one replication study was undertaken to assess female preferences for masculine male bodies across the menstrual cycle of heterosexual women. With a pre-registration, and a large sample size of 157 heterosexual women, Jünger, Kordsmeyer, Gerlach, and Penke (2018) found that while fertile women exhibited a general increase in their interest in male bodies, evaluating them as more attractive, they did not prefer more masculine men when they were more likely to be fertile. This was taken by the authors as evidence against the dual mating strategy hypothesis. In this issue, Gangestad et al. (2019) offer a critique of Jünger et al. (2018), and a reanalysis of the primary data from that study. In particular Gangestad et al. (2019) critique the use of cycle phase estimates made from counting back from menses and the use of urinary Luteinizing Hormone (LH) tests to infer the women's fertile period, and instead focus on analysis of women's preferences with respect to circulating steroid hormone concentrations, in particularly, concentrations of estrogens (E) and progesterone (P). They also critique the measures of male body masculinity used in Jünger et al. (2018). Gangestad et al. (2019) reanalyze the data provided by Jünger et al. (2018), and suggest that when using a different measure of masculinity, multi-level regression reveals an interaction effect between the E:P ratio, relationship status, and mate preference. In responding to this critique, Stern, Arslan, Gerlach, and Penke (2019) take a multiverse analytic approach and argue that the effects reported by Gangestad

E-mail address: jhigham@nyu.edu.

https://doi.org/10.1016/j.evolhumbehav.2019.08.007
Received 13 August 2019; Received in revised form 22 August 2019; Accepted 24 August 2019

et al. (2019) are not robust, and are only found with one particular set of analytical decisions.

2. Measuring menstrual cycle phase

One discussion that is central to the debate is what the best measures of underlying menstrual cycle variation might be. Jünger et al. (2018) estimated the timing of the cycle phase using the reverse cycle day method based on the estimated day of the next menstrual onset, with the fertile phase then confirmed by the use of LH strip measurement. A prior simulation study found that this method was the most accurate currently being employed to estimate conceptive probability, and recommended this as the preferred method (Gangestad et al., 2016). In addition, that simulation study also recommended that measures of steroid hormones be taken in future studies (Gangestad et al., 2016). In their present commentary, Gangestad et al. (2019) propose that the best measures of cycle phase to use are based on concentrations of E and P (assessed in their pre-registered study by the measurement of E1C and PdG, respectively), and in particular, the log of the ratio between them, ln(E:P). Since estrogen concentrations rise in the follicular phase of the cycle, peaking around the peri-ovulatory phase before falling after ovulation, and since progesterone concentrations rise quickly after ovulation following the formation of the corpus luteum, the relative ratio of the concentration of these hormones is a good indicator of the cycle phase, and the switch to lower ratios after ovulation is sometimes referred to as the Day of Luteal Transition (DLT). So, which of these methods – the estimation of conceptive probability using the counting method combined with LH measurement, versus the measurement of the E:P ratio, is the most appropriate method to use?

Here, comparative work may offer some insight. For those of us that commonly ask questions about the relationships between the cycle phase, female signals of ovulation, and female mate choice in species of nonhuman primate such as baboons, macaques, and chimpanzees, we often only have estimates of E and P concentrations to infer the female's underlying cycle phase. This is because we are typically reliant on measuring E and P metabolite concentrations from urine or fecal samples. Here, the measurement of LH is often problematic, partly because the form of LH seems to differ outside of the apes, and partly because very frequent urine sampling would be necessary to ensure that the LH peak was correctly identified. Such studies often analyze changes in female signals, female behavior, and male behavior, with respect to estimates of the fertile phase made from E and P hormone metabolite concentrations (e.g. chimpanzees, Emery & Whitten, 2003, Deschner, Heistermann, Hodges, & Boesch, 2003, 2004; bonobos, Douglas, Hohmann, Murtagh, Thiessen-Bock, & Deschner, 2016; white-handed gibbons, Barelli, Heistermann, Boesch, & Reichard, 2007; long-tailed macaques, Engelhardt, Hodges, Niemitz, & Heistermann, 2005; Engelhardt et al., 2004; Barbary macaques, Brauch et al., 2007, Pfefferle, Brauch, Heistermann, Hodges, & Fischer, 2007; rhesus macaques, Dubuc et al., 2009, Dubuc, Muniz, Heistermann, Widdig, & Engelhardt, 2012; crested macaques, Higham et al., 2012; olive baboons, Higham, MacLarnon, Ross, Heistermann, & Semple, 2008; Higham, Semple, MacLarnon, Heistermann, & Ross, 2009). In addition to assessing whether signals and behavior change with respect to these estimates, many studies have also *separately* asked the question of whether signals and behavior change with respect to the hormone metabolite concentrations themselves (e.g. long-tailed macaques, Engelhardt et al., 2005; Barbary macaques, Pfefferle, Heistermann, Pirow, Hodges, & Fischer, 2011; crested macaques, Higham et al., 2012). To create these separate measures from the same underlying E and/or P data, each individual cycle profile is assessed separately. For example, a change in hormone concentrations greater than 2 standard deviations above an established mean baseline, and sustained for several consecutive values, can be used to indicate a significant rise (Jeffcoate, 1983). Detecting such a change in P concentrations can be used to determine the onset of the luteal phase. When such methods are applied to excreta, it is also important to incorporate excretion lags that delay how concentrations of native hormones in blood are reflected by hormone metabolite excretion into urine or feces (e.g. Wasser, Monfort, Southers, & Wildt, 1994). Once such estimates are created, one set of questions can be asked about whether signals and behavior relate to estimates of the timing of the fertile phase and conceptive probability, which are important for addressing questions about the evolutionary function of changes in signals and behavior. Separate questions can also be asked about whether this variation relates to the underlying E and P concentrations themselves. One emergent result from studies that have asked such questions, is that there is variation in how female signals and behavior are connected to the timing of the fertile phase, and separately, how they respond to underlying changes in E and P across the cycle. This variation occurs at three levels: inter-specific, inter-individual, and intra-individual inter-cycle.

3. Variation in the effects of steroid hormone concentrations on primate signals and behavior

In species of non-human primate, there are often multiple changes seen across the cycle, that all appear to be collectively linked to changes in concentrations of E and P, and hence to the timing of ovulation. The general conceptual template for ovarian signaling and changes in primates is that, for those species exhibiting these signals and behavior, during the follicular phase of the cycle as estrogen concentrations rise, sexual swellings inflate, the probability of copulation calls being given increases, and female behavioral attractiveness, receptivity, and proceptivity all increase. Following ovulation, and the concomitant drop in circulating E and rise in post-ovulatory P, these signals and behaviors decrease in expression. These effects can be induced experimentally. For example, estrogen injections given to baboons stimulate tumescence of the anogenital swelling (Parkes & Zuckerman, 1931), while swelling detumescence can be elicited by progesterone injections (Gillmann, 1940). Similarly, the administration of estradiol to ovariectomized rhesus macaques causes an increase in female sexual attractiveness, proceptivity, and receptivity (Johnson & Phoenix, 1976; Michael & Saayman, 1968), and these effects can be inhibited by injections of progesterone (Michael, Saayman, & Zumpe, 1968). Similar experimental effects have been observed in chacma baboons (Saayman, 1968). Effects of estrogen and progesterone are thought to be particularly strong on female attractiveness in rhesus macaques (Baum, Keverne, Everitt, Herbert, & De Greef, 1977), with adrenal androgens also involved in mediating receptivity and proceptivity (Baum, Everitt, Herbert, & Keverne, 1977). In theory then, changes in multiple types of signals and behavior related to mating biology may be linked concurrently to changes in the underlying E:P ratio. However, empirical data from the field have suggested that there is in practice a great deal of variation in how such aspects of biology respond to changes in E:P.

Firstly, the extent to which different types of signals, such as sexual swelling size and proceptive behavior, are linked to changes in the concentration of these underlying hormones, and to the timing of the fertile phase show *inter-specific* variation. In some species, proceptive behavior towards males seems very tightly linked to both E:P and, separately, the timing of the fertile phase as assessed from profiles of E and P, while swelling expression is linked to E:P, but not to fertility phase timing (e.g., long-tailed macaques, Engelhardt et al., 2005). In others, expression of the swelling itself seems tightly linked to both E:P and the timing of the fertile phase as assessed from profiles of E and P, while proceptive behavior towards males is not linked to the timing of the fertile phase, and is only weakly linked to E:P (e.g., olive baboons, Higham et al., 2008, 2009). In other species still, both swelling expression and proceptive behavior towards males seem tightly linked both to each other, to the E:P ratio, and to the timing of the fertile phase (e.g., crested macaques, Higham et al., 2012). In Barbary macaques, copulation call structures are well predicted by E (positively) and P (negatively) concentrations (Pfefferle et al., 2011), but nonetheless do not indicate the timing of the fertile phase (Pfefferle et al., 2007). This discrepancy is because, while there is a general strong association between copulation call parameters and E and P concentrations, acoustic parameters do not track fine-scale changes in E in the late follicular phase that occur around the timing of ovulation (Pfefferle et al., 2011). As such, despite the fact that the proximate mechanisms regulating signals and behavior are in theory the same, the relative association of signals such as sexual swellings, copulation calls, and proceptive behavior with E:P, and estimates of the fertile phase based on E and P concentrations, seems to vary between species.

The responsiveness of specific E:P linked signals also shows variation between cycles within the same species, both at the *inter-individual*, and *intra-individual inter-cycle* level. For example, Deschner et al. (2003) measured concentrations of E and P metabolites measured from urine, and used the latter to estimate the timing of ovulation with respect to swelling detumescence. These data showed that, in some cycles, swelling detumescence occurred the day after ovulation, whereas in others it did not occur until as late as 7 days after ovulation. This variation occurred across cycles both between and within females. For example, Deschner et al. (2003) presented data for one wild chimpanzee female ("Duna") for whom 6 cycles were measured. Across these 6 cycles, detumescence occurred the day after ovulation in one cycle, 2 days after ovulation in 3 cycles, 3 days after ovulation in one cycle, and 4 days after ovulation in one cycle. The principal that swelling expression can vary with respect to both the cycle phase and hormonal changes across the cycle is in fact central to the prevailing hypothesis for how such signals function – that swellings offer only a probabilistic indication of the likelihood of ovulation, with this varying across cycles and individuals (Nunn, 1999). If swellings always responded in exactly the same way to hormonal changes across the cycle, then detumescence would be a a highly accurate indicator of ovulation.

One potential mechanism by which different signals, and behavior, could be made more or less sensitive to changes in the E:P ratio, is by tissue-specific receptor expression. Mechanisms could potentially operate that alter the expression of hormone receptors in a specific tissue, or in the brain, such that one signal or aspects of behavior become relatively insensitive to changes in hormone concentrations (Higham, 2016; Higham, Pfefferle, Heistermann, Maestripieri, & Stevens, 2013). For example, reduced progesterone-receptor expression in the tissue of a sexual swelling following ovulation would make it relatively insensitive to the post-ovulatory progesterone rise, hence maintaining swelling tumescence after ovulation.

Interestingly, an analogous effect is observed in the signals of male primates exhibiting steroid-hormone linked signals. Here, injections of testosterone (T) are known to cause reddening of bare-skin signals in the males of multiple species, including hamadryas baboons, drills (Zuckerman & Parkes, 1939), and rhesus macaques (Rhodes et al., 1997). Experiments in rhesus macaques have shown that administration of both E and T cause skin reddening, but that administration of DHT, which is not aromatizable to estrogen, does not (Rhodes et al., 1997). Moreover, T treatment in the presence of fadrozole, which inhibits aromatization of T to E, does not cause reddening, indicating that the effects of T on male red bare-skin signals is via aromatization of T to E (Rhodes et al., 1997). Data from the field again suggest a dysregulation between how responsive fertility signals and mating behavior are to changes in underlying steroid hormone concentrations. For example, despite T concentrations being clearly linked to skin reddening in rhesus macaques, male skin redness is not correlated to T metabolite excretion in free-ranging conditions (Higham et al., 2013). Similarly, in wild gelada, T metabolite excretion rises in bachelor males before they challenge a rival male for a one-male unit (OMU) (Pappano & Beehner, 2014). However, their red chest patches do not change color. If the challenge is successful, then color subsequently changes afterwards – deposed males lose their bright coloration, while the successful challengers quickly becoming redder (Bergman, Ho, & Beehner, 2009). Such effects could be mediated by tissue-specific steroid-receptor expression, as above. Another potential mechanism relates to the form in which different types of hormones are circulating in the body. For example, T can be enzymatically reduced by 5α-reductase to 5α –DHT, which can bind to androgen receptors, but which is no longer aromatizable to estrogen. Circulating 5α –DHT could then impact male behavior without causing a concomitant change in coloration of the chest patch (see discussion in Higham et al., 2013; Higham, 2016). Similarly, for behavior, it is known that only androgens that are in aromatizable form can produce effects similar to those of estrogens on the brain (Ryan, Naftolin, Reddy, Flores, & Petro, 1972).

One further interesting finding is that in some species, similar E:P ratios to those found during the late follicular phase of the ovarian cycle, are found at some stages of gestation (e.g., chimpanzees, Wallis & Lemmon, 1986; long-tailed macaques, Engelhardt, Hodges, & Heistermann, 2007). In some species, when this occurs, females typically start to exhibit similar signals to those exhibited during ovulation, in addition to receptivity to mating and proceptivity towards males (Engelhardt et al., 2007). Whether this is a non-adaptive byproduct of gestation processes, or is a functional shift to confuse paternity is unclear, but evidence from at least some species points towards the latter (e.g. long-tailed macaques, Engelhardt et al., 2007).

4. Conclusion

Measuring cycle phase and likely conceptive probability from methods such as the count-back from menses and supported by LH strip assessment, are generally reasonably accurate methods for assessing cycle phase (Gangestad et al., 2016). Asking how signals and behavioral changes relate to cycle phase and, separately, to E:P are not necessarily the same questions. Indeed, when these two variables have been assessed separately in studies of human menstrual cycles, differing results have been found. For example, Fischer et al. (2011) found marginal changes in women's vocal pitch around ovulation, but that these parameters were not related to E or P concentrations across the cycle. While E:P does in theory determine the expression of changes in signals and behavioral proceptivity, field data support the idea that the effects of these hormones on signals and behavior show inter-specific, inter-individual, and intra-individual inter-cycle, variation. Changes in tissue-specific receptor expression, and enzymatic conversion between different steroid hormones, provide potential mechanisms by how such variation can be produced. Given the above, asking questions about variation in signals and behavior according to cycle phase measured independently of steroid hormone concentrations, or from estimates of cycle phase based on those concentrations, and about variation with respect to changes in E:P, are separately important questions.

Acknowledgements

I would like to thank Michael Heistermann and the editor, Debra Lieberman, for feedback and comments on this manuscript.

References

Barelli, C., Heistermann, M., Boesch, C., & Reichard, U. H. (2007). Sexual swellings in wild white-handed gibbon females (Hylobates lar) indicate the probability of ovulation. *Hormones and Behavior, 51*, 221–230.

Baum, M. J., Everitt, B. J., Herbert, J., & Keverne, E. B. (1977). Hormonal basis of proceptivity and receptivity in female primates. *Archives of Sexual Behavior, 6*, 173–192.

Baum, M. J., Keverne, E. B., Everitt, B. J., Herbert, J., & De Greef, W. J. (1977). Effects of progesterone and estradiol on sexual attractivity of female rhesus monkeys. *Physiology & Behavior, 18*, 659–670.

Bergman, T. J., Ho, L., & Beehner, J. C. (2009). Chest color and social status in male geladas (Theropithecus gelada). *International Journal of Primatology, 30*, 791–806.

Brauch, K., Pfefferle, D., Hodges, K., Möhle, U., Fischer, J., & Heistermann, M. (2007). Female sexual behavior and sexual swelling size as potential cues for males to discern the female fertile phase in free-ranging Barbary macaques (Macaca sylvanus) of Gibraltar. *Hormones and Behavior, 52*, 375–383.

Bryant, G. A., & Haselton, M. G. (2009). Vocal cues of ovulation in human females. *Biology*

Letters, 5, 12–15.

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review, 100*, 204–232.

Deschner, T., Heistermann, M., Hodges, K., & Boesch, C. (2003). Timing and probability of ovulation in relation to sex skin swelling in wild west African chimpanzees, pan troglodytes verus. *Animal Behaviour, 66*, 551–560.

Deschner, T., Heistermann, M., Hodges, K., & Boesch, C. (2004). Female sexual swelling size, timing of ovulation, and male behavior in wild west African chimpanzees. *Hormones and Behavior, 46*, 204–215.

Douglas, P. H., Hohmann, G., Murtagh, R., Thiessen-Bock, R., & Deschner, T. (2016). Mixed messages: Wild female bonobos show high variability in the timing of ovulation in relation to sexual swelling patterns. *BMC Evolutionary Biology, 16*, 140.

Dubuc, C., Brent, L. J., Accamando, A. K., Gerald, M. S., MacLarnon, A., Semple, S., Heistermann, M., & Engelhardt, A. (2009). Sexual skin color contains information about the timing of the fertile phase in free-ranging Macaca mulatta. *International Journal of Primatology, 30*(6), 777–789.

Dubuc, C., Muniz, L., Heistermann, M., Widdig, A., & Engelhardt, A. (2012). Do males time their mate-guarding effort with the fertile phase in order to secure fertilisation in Cayo Santiago rhesus macaques? *Hormones and Behavior, 61*, 696–705.

Emery, M. A., & Whitten, P. L. (2003). Size of sexual swellings reflects ovarian function in chimpanzees (pan troglodytes). *Behavioral Ecology and Sociobiology, 54*, 340–351.

Engelhardt, A., Hodges, J. K., & Heistermann, M. (2007). Post-conception mating in wild long-tailed macaques (Macaca fascicularis): Characterization, endocrine correlates and functional significance. *Hormones and Behavior, 51*, 3–10.

Engelhardt, A., Hodges, J. K., Niemitz, C., & Heistermann, M. (2005). Female sexual behavior, but not sex skin swelling, reliably indicates the timing of the fertile phase in wild long-tailed macaques (Macaca fascicularis). *Hormones and Behavior, 47*, 195–204.

Engelhardt, A., Pfeifer, J. B., Heistermann, M., Niemitz, C., van Hooff, J. A., & Hodges, J. K. (2004). Assessment of female reproductive status by male longtailed macaques, Macaca fascicularis, under natural conditions. *Animal Behaviour, 67*, 915–924.

Fischer, J., Semple, S., Fickenscher, G., Jürgens, R., Kruse, E., Heistermann, M., & Amir, O. (2011). Do women's voices provide cues of the likelihood of ovulation? The importance of sampling regime. *PLoS ONE, 6*, e24490.

Gangestad, S. W., Dinh, T., Grebe, N. M., Del Giudice, M., & Emery Thompson, M. (2019). Psychological cycle shifts redux: Revisiting a preregistered study examining preferences for muscularity. *Evolution and Human Behavior* (This issue).

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*, 151–163.

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.

Gangestad, S. W., & Simpson, J. A. (2000). The evolution of human mating: Trade-offs and strategic pluralism. *The Behavioral and Brain Sciences, 23*, 573–587.

Gangestad, S. W., Simpson, J. A., Cousins, A. J., Garver-Apgar, C. E., & Christensen, P. N. (2004). Women’s preferences for male behavioral displays change across the menstrual cycle. *Psychological Science, 15*, 203–207.

Gangestad, S. W., & Thornhill, R. (1998). Menstrual cycle variation in women’s preferences for the scent of symmetrical men. *Proceedings of the Royal Society B, 265*, 927–933.

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*, 1205–1259.

Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014b). Meta-analyses and p-curves support robust cycle shifts in women’s mate preferences: Reply to Wood and Carden (2014) and Harris, Pashler, and Mickes (2014). *Psychological Bulletin, 140*, 1272–1280.

Gillmann, J. (1940). The effect of multiple injections of progesterone on the turgescent perineum of the baboon (Papio porcarius). *Endocrinology, 26*, 1072–1107.

Haselton, M. G., Mortezaie, M., Pillsworth, E. G., Bleske-Rechek, A., & Frederick, D. A. (2007). Ovulatory shifts in human female ornamentation: Near ovulation, women dress to impress. *Hormones and Behavior, 51*, 40–45.

Havlíček, J., Roberts, S. C., & Flegr, J. (2005). Women’s preference for dominant male odour: Effects of menstrual cycle and relationship status. *Biology Letters, 1*, 256–259.

Higham, J. P. (2016). Field endocrinology of nonhuman primates: Past, present and future. *Hormones and Behavior, 84*, 145–155.

Higham, J. P., Heistermann, M., Saggau, C., Agil, M., Perwitasari-Farajallah, D., & Engelhardt, A. (2012). Sexual signaling in the crested macaque and the evolution of primate fertility signals. *BMC Evolutionary Biology, 12*, 89.

Higham, J. P., MacLarnon, A., Ross, C., Heistermann, M., & Semple, S. (2008). Baboon sexual swellings: Information content of size and color. *Hormones and Behavior, 53*, 452–462.

Higham, J. P., Pfefferle, D., Heistermann, M., Maestripieri, D., & Stevens, M. (2013). Signaling in multiple modalities in male rhesus macaques: Barks and sex skin coloration in relation to androgen levels, social status and mating behavior. *Behavioral Ecology and Sociobiology, 67*, 1457–1469.

Higham, J. P., Semple, S., MacLarnon, A., Heistermann, M., & Ross, C. (2009). Female reproductive signals, and male mating behavior, in the olive baboon. *Hormones and Behavior, 55*, 60–67.

Jeffcoate, S. L. (1983). Use of rapid hormone assays in the prediction of ovulation. In S. L. Jeffcoate (Ed.). *Ovulation: Methods for its prediction and detection* (pp. 67–82). Chichester: John Wiley & Sons Ltd.

Johnson, D. F., & Phoenix, C. H. (1976). Hormonal control of female sexual attractiveness, proceptivity, and receptivity in rhesus monkeys. *Journal of Comparative and Physiological Psychology, 90*, 473–483.

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.

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.

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

Michael, R. P., & Saayman, G. S. (1968). Differential effects on behaviour of the subcutaneous and intravaginal administration of oestrogen in the rhesus monkey (Macaca mulatta). *The Journal of Endocrinology, 41*, 231–246.

Michael, R. P., Saayman, G. S., & Zumpe, D. (1968). The suppression of mounting behaviour and ejaculation in male rhesus monkeys (Macaca mulatta) by administration of progesterone to their female partners. *The Journal of Endocrinology, 41*, 421–431.

Nunn, C. L. (1999). The evolution of exaggerated sexual swellings inprimates and the graded-signal hypothesis. *Animal Behaviour, 58*, 229–246.

Pappano, D. J., & Beehner, J. C. (2014). Harem-holding males do not rise to the challenge: Androgens respond to social but not to seasonal challenges in wild geladas. *Royal Society Open Science, 1*, 140081.

Parkes, A. S., & Zuckerman, S. (1931). Some effects of oestrin on the menstrual cycle of baboons and macaques. *Journal of Anatomy, 65*, 272.

Pashler, H., & Wagenmakers, E. J. (2012). Editors’ introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528–530.

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, 39*, 741–742.

Pfefferle, D., Brauch, K., Heistermann, M., Hodges, J. K., & Fischer, J. (2007). Female Barbary macaque (Macaca sylvanus) copulation calls do not reveal the fertile phase but influence mating outcome. *Proc. Roy. Soc. B. 275*, 571–578.

Pfefferle, D., Heistermann, M., Pirow, R., Hodges, J. K., & Fischer, J. (2011). Estrogen and progestogen correlates of the structure of female copulation calls in semi-free-ranging Barbary macaques (Macaca sylvanus). *International Journal of Primatology, 32*, 992–1006.

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

Rhodes, L., Argersinger, M. E., Gantert, L. T., Friscino, B. H., Hom, G., Pikounis, B., ... Rhodes, W. L. (1997). Effects of administration of testosterone, dihydrotestosterone, oestrogen and fadrozole, an aromatase inhibitor, on sex skin colour in intact male rhesus macaques. *Reproduction, 111*, 51–57.

Ryan, K. J., Naftolin, F., Reddy, V., Flores, F., & Petro, Z. (1972). Estrogen formation in the brain. *American Journal of Obstetrics and Gynecology, 114*, 454–460.

Saayman, G. S. (1968). Oestrogen, behaviour and permeability of a troop of chacma baboons. *Nature, 220*, 1339.

Stern, J., Arslan, R. C., Gerlach, T. M., & Penke, L. (2019). No robust evidence for cycle shifts in preferences for men's bodies in a multiverse analysis: A response to Gangestad et al. (2019). *Evolution and Human Behavior* (This issue).

Thornhill, R., Gangestad, S. W., Miller, R., Scheyd, G., McCollough, J. K., & Franklin, M. (2003). Major histocompatibility complex genes, symmetry, and body scent attractiveness in men and women. *Behavioral Ecology, 14*(5), 668–678.

Wallis, J., & Lemmon, W. B. (1986). Social behavior and genital swelling in pregnant chimpanzees (pan troglodytes). *American Journal of Primatology, 10*, 171–183.

Wasser, S. K., Monfort, S. L., Southers, J., & Wildt, D. E. (1994). Excretion rates and metabolites of oestradiol and progesterone in baboon (Papio cynocephalus cynocephalus) faeces. *Journal of Reproduction and Fertility, 101*, 213–220.

Wood, W., & Carden, L. (2014). Elusiveness of menstrual cycle effects on mate preferences: Comment on Gildersleeve, Haselton, and Fales (2014). *Psychological Bulletin, 140*, 1265–1271.

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.

Zuckerman, S., & Parkes, A. S. (1939). Observations on the secondary sexual characteristics in monkeys. *J.Endocrinol. 1*, 430–439.

Evolution and Human Behavior 40 (2019) 537–542

Contents lists available at ScienceDirect

Evolution and Human Behavior

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

Commentary

Psychological cycle shifts redux, once again: response to Stern et al., Roney, Jones et al., and Higham

Steven W. Gangestad[a,*], Tran Dinh[a], Nicholas M. Grebe[b], Marco Del Giudice[a], Melissa Emery Thompson[c]

[a] Department of Psychology, University of New Mexico, United States of America
[b] Department of Evolutionary Anthropology, Duke University, United States of America
[c] Department of Anthropology, University of New Mexico, United States of America

Our target article presented a critical reanalysis of an impressive dataset published by Jünger et al. on cycle shift differences. Jünger, Kordsmeyer, Gerlach, and Penke (2018) had made bold, definitive claims: Cycle shifts "do not seem to alter preferences for body characteristics at all, leaving no room for cycle shifts in mate preferences for masculine characteristics or any other assumed indicators of good genes" (p. 421). Our article had three goals. First, we reanalyzed their publicly-available data to examine if their null finding was robust to modest differences in approach. Second, we sought to determine—and indeed found that—the portions of Jüngers et al.'s preregistration that were omitted from their analyses affected their conclusion. Third, we sought to provide some productive discussion on the advantages and limitations of preregistration. The commentaries speak to specific aspects of our claims and the evidence for them, as well as broader issues regarding scientific inquiry: strategies for scientific progress, exploratory analysis, secondary data analysis.

In this brief response, we address several major issues raised by commentators. Our response is organized into 7 sections, the titles of which state our primary claims.

Before getting into these matters, however, we note two points of agreement with Jünger et al. (now Stern et al., this issue). Their null assertion partly motivated our target article. Even our modest message that effects *may* exist represents a sharp contrast against the background of a strong null assertion. Relatedly, we stressed the general point that, while preregistration obligates scholars to proceed with a particular analysis, data from the preregistered study itself can call into question interpretations from that analysis. Stern et al. "agree that one should not make strong conclusions in favor of the null hypothesis too early, especially not based on a single study" (p. XXX), even one that is preregistered.

1. Stern et al.'s multiverse analysis betrays the logic of multiverse analysis and does not support their claims

Stern et al.'s commentary culminates in a multiverse analysis, which they claim "provides evidence that [our] results are not robust." In our view, Stern et al.'s multiverse analysis cannot show what they claim because it severely deviates from the logic of multiverse analysis.

The idea of a multiverse stems from the notion that many possible analyses testing a particular effect can be constructed from a data set. Even when researchers explore just one "forking path" analytically, there may be many equally justifiable paths (e.g., Gelman & Loken, 2014), producing a "multiverse" of results. Steegen, Tuerlinckx, Gelman, and Vanpaemel (2016) proposed doing analyses all possible ways within a multiverse when, in fact, choices are arbitrary, "whimsical," and lack clear justification.

The multiverse notion is an important one. But of course, it also implies a specific domain of appropriate applicability. Naturally, neither Gelman and Loken (2014) nor Steegen et al. (2016) argue against researchers choosing justified analyses over unjustified ones; that would be silly. In an appropriate multiverse analysis, then, one does not evaluate the robustness of results from justified analyses by asking how they compare to results from unjustifiable, poor ones. As Steegen et al. (2016) explicitly state, "This practice of selective reporting *would not be problematic if the single data set under consideration is processed based on sound and justifiable choices*" (p. 703; emphasis added). Rather, they propose that one explore a multiverse defined by a set of alternatives that are *equally* justifiable.

There are many ways to generate, from a decision tree, a collection of weak, unjustifiable tests. An invalid measure may substitute for a valid one. A valid measure can be split into unreliable components. Or, one can enter multiple correlated valid indicators of the same trait simultaneously, which fractionates valid variance captured by each. Consider an example. Suppose the predictor of a criterion is a personality trait. This trait has been assessed with 10 items, but only 5 turn out to be valid. The best measure of the trait is a composite of those 5. But of course, one can generate many alternative models: e.g., using the total sum of 10 entered; entering items separately; entering items simultaneously. In the latter analysis, even each valid item likely has very little residual validity, as its valid variance overlaps with that of the

[*] Corresponding author at: Department of Psychology, University of New Mexico, Albuquerque, NM 87131, United States of America.
E-mail address: sgangest@unm.edu (S.W. Gangestad).

https://doi.org/10.1016/j.evolhumbehav.2019.08.008
Received 25 August 2019; Accepted 27 August 2019

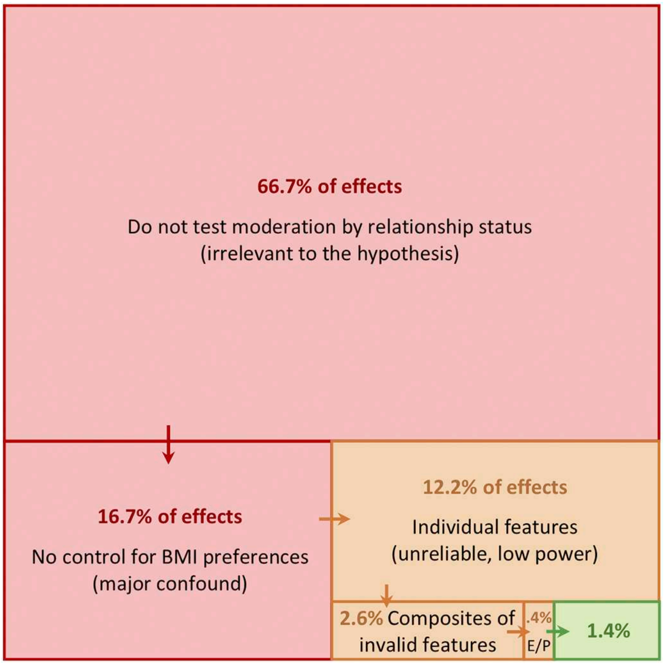

Fig. 1. The distribution of effects in Stern et al.'s multiverse. Starting at the top, 66.7% of effects concern a hypothesized effect separate from the moderation effect; separate conceptual effects demand separate multiverse analyses. Half of the 33.3% that remain—16.7% of the total—do not control for BMI preferences, which Stern et al. agree should be controlled. Within the 16.7% that now remain, the vast majority—12.2% of the total—use unreliable single item indicators. Of the remaining composites, just 1.8% aggregate items that pass basic validation tests. Sollberger and Ehlert (2016) warn against using raw hormone ratios, and the E/P ratio in Jünger et al.'s data does not straightforwardly tap additive or even simple interactive E and P effect. Remaining, reasonably justifiable effects constitute 1.4% of the 1254 effects that Stern et al. include.

remaining 4, which is partialled out. The combinatorial nature of decisions means that many weakly powered tests can be generated from few decisions. It would be nonsensical to argue that an effect found with the optimal composite is "not robust" because its effects are dwarfed by those of very weak alternatives.

Stern et al. generate 416 models with 1254 effects, drawing a distinction between this dazzling number of irrelevant or unreliable effects and the relatively small number that we test. Yet their analysis goes against a number of explicit recommendations mentioned above. See Fig. 1. First, they confound effects. The 3-way relationship status moderation effect and the 2-way interactions not involving moderation with relationship status are different effects. Steegen et al. (2016) sensibly treat different effects as distinct (e.g., their Fig. 1). In Stern et al.'s multiverse, by contrast, half of the effects in models with relationship status are irrelevant two-way interactions, and half of the models do not include relationship status as a factor at all (which Stern et al. agree should be included) and, hence, only have 2-way interactions. As a result, only 33% of effects constitute the relationship moderation effect of theoretical interest. Second, of these, half do not control for BMI preferences, which Stern et al. agree is clearly preferable; that leaves 17% of effects.

The remaining effects involve different ways to specify the 7 male features that Jünger et al. used to operationalize cues of male upper body strength or formidability. In our target article, we present critiques of the validity of these measures and how they were entered into the models. Jünger et al.“s own stimulus dataset indicates that 5 out of 7 measures fail the basic validation of predicting either sexual attractiveness or bodily dominance. Of the moderation effects with BMI controlled, then, the large majority (74%) involve single 'item' predictors (half the time controlling for all other items), the vast majority of which did not pass validation tests. The small proportion of effects that involve composites (now down to 4%) include two factors that effectively aggregate only invalid features and Stern et al.”s own composite (see below for further discussion). The remaining composites (< 2%) include the two validated features, or a broader factor reflecting their common dimension that we constructed, to most sensitively assess visual cues of upper body strength. Therefore, in Stern et al.'s multiverse, effects from these composites are overwhelmed by a sea of effects that are either entirely irrelevant or possess miniscule power to detect effects of interest. *Even if* true moderation on hormonal association with preferences for muscularity exists, the only reasonable expectation is largely a set of null effects. That pattern therefore cannot be diagnostic of the effect being "not robust."

We add one other consideration. In Jünger et al.'s data, there exist two massive outliers (0.3%) on raw progesterone levels (belonging to the same participant), 8 and 22 standard deviations above the mean of all remaining values. The distance of these outliers from the remaining 99.7% data points is > 2× and 5× the full range, top to bottom, of that 99.7% (see Fig. S1, Supplementary Online Materials [SOM]). For their multiverse analysis, Stern et al. included those outliers without informing readers. The outliers massively leverage outcomes. Consistent with other analyses (e.g., Jones et al., 2018; Roney & Simmons, 2013), we removed these outliers.

Fig. 3 shows results (with the two outliers removed) in a set of analyses that are reasonably justified. In both their set and our broader set, results are not inconsistent with true relationship status moderation effects on P associations.[1]

Stern et al.'s multiverse analysis has broader implications. Multiverse analysis can be a valuable tool when decisions are truly arbitrary or equally defensible. But Stern et al.'s multiverse illustrates its hazards, as the approach is vulnerable to the proliferation of poor-specified models producing null results. Its appropriate use requires scrutiny of the justifiability of effects within the multiverse.

In their commentary, Jones et al. explicitly observe that secondary data analyses can be valuable, but note the risks of drawing conclusions from them, given that analysis plans may be informed by the very data analyzed. In a constructive spirit, they describe four strategies to increase the trustworthiness of such analyses, one of which is multiverse analysis (or, relatedly, specification curve analysis). We agree with the sentiment and report 38 different analyses ourselves (Table 8, target article). Jones et al., however, uncritically accept Stern et al.'s multiverse analysis. We think this illustrates the danger we describe: Now, not one, but two papers recommend a multiverse analysis that contradicts the method's foundations.

2. Stern et al. and Jones et al. ignore hormonal associations with preferences for bodily dominance, a crucial set of findings

2.1. Bodily dominance effects do not support stern et al.'s explanation for our effects

Stern et al. paint our findings as highly selective and thus not robust. However, they ignore a critical set of our analyses. We evaluated the validity of cues of male muscularity by examining their association with independent observers' ratings of Bodily Dominance (formidability), which correlate highly with bodily attractiveness. Though not a perfect criterion of muscularity, Bodily Dominance (with BMI controlled) is likely a better measure than any single physical cue, and likely a better

[1] We stress that these analyses come from Stern et al.'s multiverse, not the one we would generate. In our preregistration, we enter between-woman hormone values as separate predictors, as these too may account for variance, and as recommended by West, Ryu, Kwak, and Chan (2011). Addition of these effects—which we argue is clearly justified—partly explains why the 38 effects we present are all significant.

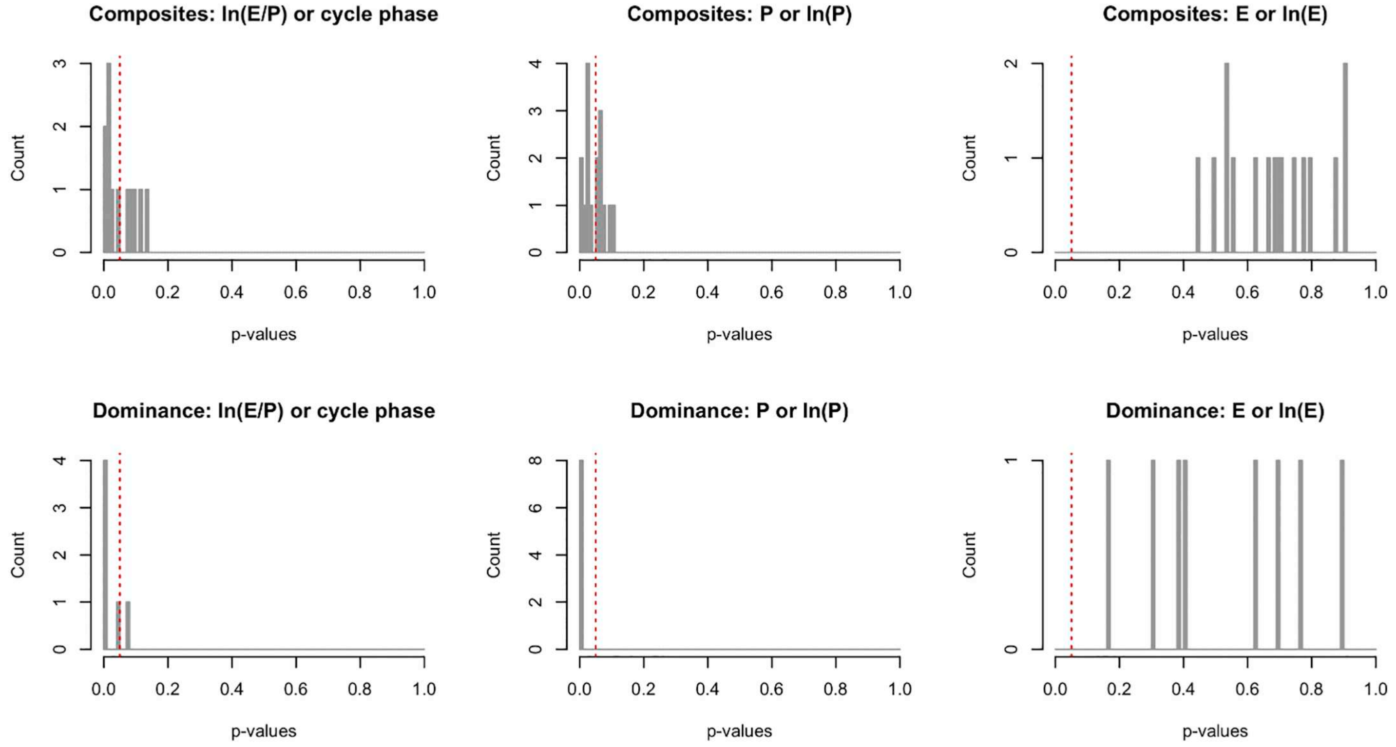

Fig. 2. Distribution of *p*-values in a multiverse of effects using body feature composites (top panels) and Bodily Dominance ratings (bottom panels). Two composites are used: our empirically vetted composite of Strength/Muscularity; and a factor tapping this dimension extracted from all 7 male features. (We used our previously reported factor scores for these analyses; see target article.) We included both total and within-woman mean-centered hormone values for ln(E/P), ln(P), ln(E), raw P and raw E. Analyses both controlling for testosterone levels and not controlling for testosterone levels (target article) are included. Table S1, Supplementary Online Materials, reports p-values for all individual effects. All analyses in R included in SOM.

measure than our composite of two cues. From viewing 3D bodies, raters could perceive more than just a few features when judging Bodily Dominance. It follows that analyses should show similar or stronger results when substituting our muscularity composite with Bodily Dominance—and they did. In parallel multiverse analyses (Fig. 2), all *p* for hormonal analyses were ≤ 0.008; *p* for cycle phase were 0.072 (full sample) and 0.048 (preregistered sample of 112).

Stern et al. claim that our composite was overfitted through particular attention to two cues. Though we justified our selection of these two cues—they were the only ones that met straightforward validation criteria—one might wonder whether we sifted through many combinations of the 7 features, landed on 2 that worked, and thereby capitalized on chance error. But Bodily Dominance was not one of these 7 features; indeed, it was not considered by Jünger et al. at all. It stands apart as a singular rating that unassailably relates to muscularity. One can then wonder how Stern et al.'s view explains the fact that Bodily Dominance generates strong moderation effects. In their view, it must have nothing whatsoever to do with preferences for muscularity. Neither Stern et al. nor Jones et al. (who endorse Sterns et al.'s view) address these findings. The questions are simple: How does capitalization on chance error in our analyses using our composite measure explain strong moderation effects for Bodily Dominance? And, if capitalization on chance errors cannot explain findings for Bodily Dominance, what is the likelihood that our findings for Strength/Muscularity are explained, as Stern et al. imply, by capitalization on chance errors?

2.2. We performed additional analyses for robustness checks

We did not present a single analysis. We used multiple measures of male muscularity, and hormone values were treated as both logged and raw values. Table 8 (target article) presents 38 alternative analyses involving ln(E/P), ln(P), or raw P.

3. Stern et al. reveal that Jünger et al.'s composite measure, though flawed, did yield relationship status moderation of cycle effects; they should have reported this finding

Stern et al. reveal that they constructed a composite measure of all 7 variables themselves, which “did not change results” (though this analysis was not reported in Jünger et al.'s paper). The issue is the same as the one we identified in our target article: If 5 of 7 features bear little to no association with muscularity, the aggregate is likely to have only modest validity. (Indeed, as we show in Table 2, target article, the correlation of shoulder-to-chest ratio with Bodily Dominance without BMI controlled is *negative* [−0.37]; adding this feature to a composite can greatly weaken its validity.)

Results based on this composite, nevertheless, speak to Stern et al.'s claims about their findings. In addition to claiming that analyses based on this composite did not change results, they wrote that they similarly did not report preregistered analyses examining moderation by relationship status, partly because they “led to unaltered conclusions.” They cite their Table S5, which pits all 7 predictors against one another, a procedure that weakens power greatly. In their results in Table S6, however, they report *significant* moderation effects involving their composite measure, both controlling and without controlling for BMI ($p = .043$ and 0.049.) Hence, even in their own analyses, these authors find some evidence for moderation—evidence that went completely unreported in Jünger et al. (2018). Jünger et al. (2018) should have reported moderation effects, significant or not. But had Jünger et al. (2018) reported these significant effects, readers would have an altered sense of their findings. Although the findings are not definitive evidence for moderation, they do not warrant strong null assertions.[2,3]

4. We followed our own preregistration, but fully acknowledged that aspects of our re-analyses of Jünger et al.'s data were data-dependent

As noted previously, we preregistered a study that examines hormonal associations with preferences. The preregistration, based on a funded NSF grant proposal, was initially submitted on February 21, 2018 for journal review. A revision was submitted March 11, 2018. Final acceptance was not received until later, such that it was posted on Open Science Framework April 18, 2018. As we openly stated, this preregistration was not designed to reanalyze Jünger et al.'s data. To be consistent with our past and future planned analyses, we imported several features regarding z of the preregistration, drafted well before we downloaded Jünger et al.'s data (March 17, 2018) or received ratings of Bodily Dominance from Tobias Kordsmeyer (April 6, 2018). Of course, we could not possibly have preregistered aspects of Jünger et al.'s design differing from our own, notably concerning male stimuli. Our analyses were data-dependent (though see section on Bodily Dominance above on ways we attempted to address data-dependence). Stern et al. assert, “contrary to their claim, the exact analyses they did were never preregistered by anyone.” But of course we never said anything to the contrary.

5. Stern et al. mischaracterize Marcinkowska, Kaminski, Little, and Jasienska's (2018) finding, which runs in the same direction as the finding we report from Jünger et al.

After we found evidence for moderation in Jünger et al.'s (2018) data, we learned that Marcinkowska et al. (2018) reported similar moderation of P associations with preferences for masculine bodies. They found a positive association between P and preferences for single women, and a non-significant negative (near-zero) association for partnered women (though power to detect effects within either group is low). In Jünger et al.'s data, we too found a more positive association between P and preferences for singles than partnered women (see Table 6, last line).[4] Contrary to Stern et al.'s claims, then, these moderation effects run in the same direction.

6. Control by steroid hormones is the only coherent theory of how behavioral shifts arise physiologically, though the precise mechanisms of hormonal control remain imperfectly understood

6.1. *Cycle shifts reflect coordinating effects of steroid hormones, which, functionally and physiologically, need not perfectly track conception status*

Functionally, evolutionary frameworks concerning cycle shifts highlight fecundability (conceptive status), as it varies across the cycle. Conceptive status, in turn, depends on temporal proximity to ovulation (though, importantly, ovulatory cycles are not equally fecund). At a proximate level, however, cycle shifts occur through specific mechanisms. Levels of steroid hormone, notably E and P, shift across the cycle, which functions to coordinate activities across multiple physiological systems. Indeed, there exists no alternative coherent theory about how cycle phase shifts arise. Naturally, then, a physiological focus on conceptive status *implies* a focus on hormonal effects (cf. Roney).

Higham brings comparative primate data to bear on the question how E and P levels associate with conceptive status. These data reveal both intra- and inter-specific variation in how E and P relate to conceptive status. Higham concludes that examination of how hormones and conceptive status relate to female physiological and psychological features are at least slightly different questions (though, given the function of hormonal systems, they cannot be treated as unrelated). These observations are both interesting and valuable. From a functional standpoint, a key variable is conception risk. But, as we discussed in our SOM (target article), for good reason adaptive effects may not perfectly track conception risk (cf. Roney). For instance, adaptive behavior in the early follicular phase, prior to ovulation, may differ from that in the luteal phase, after ovulation, despite both phases being associated with low conception risk.

As Roney describes, the effects of E and P are not merely immediate. They stimulate proliferation of receptors, with temporally downstream effects, and genomic effects may be delayed (see, e.g., Roney & Simmons, 2013). These effects introduce challenges to empirical study of E and P effects. As well, sampling during a conceptive phase in a study is generally highly diverse physiologically—in Jünger et al.'s study, up to a week apart, relative to the LH surge.

6.2. *The moderation effects we report do not depend on log-transformation*

Roney questions the validity of using log-transformed E and P measures. The issues he discusses are potentially important, though complex. The most important point for this response is that, in fact, the findings we presented in the target article do not depend on log-transformation. We reported analyses with both logged and raw values, finding interactions with both (see, e.g., Table 8, target article). Stern et al. claim to not find significant associations with raw hormone values, contrary to our claims. But as we already noted, they retained two extreme outliers on P, 8 and 22 standard deviations apart from all other values; we removed these two outliers. Logging P brings outliers much closer to the core distribution, merely ~0.5 and ~1 standard deviations from all other values (see Fig. S1, SOM); ln(P) results are hence not as severely affected by outliers. In multiverse analyses presented in Fig. 3, with the two outliers removed, moderation effects using raw or logged P do not meaningfully differ. Indeed, in analyses on Bodily Dominance, the mean *p*-values when raw progesterone values are used are lower than those when ln(P) values are used (0.002 vs. 004). The claim that effects crucially depend on log-transformation is simply not true.

We offer a few reflections on Roney's claims about raw vs. log-transformed hormone ratios. (a) Though the E/P ratio may track conception risk very well, that need not imply that the physiological and behavioral effects of E and P are captured by the ratio. The ratio peaks near maximal conception risk a day or two prior to ovulation because of its temporal association with the event of ovulation itself, not because peak E/P exerts immediate causal effects on it (or, for that matter, adaptive behavior). As Higham notes and we discuss above, the effects of E and P need not perfectly track conception risk. (b) The fact that the E/P ratio reflects complex non-linear interactions between E and P offers no assurance that the ratio *appropriately* captures true *E*-P interaction effects. As Sollberger and Ehlert (2016) advise, researchers should model interaction effects (e.g., by inclusion of E × P terms) rather than blindly assume that a hormone ratio captures interaction effects. (c) A widely-adapted model argues that ligands' binding affinities are a sigmoidal function of the log of the availability of the ligand (e.g., hormone concentration), with some justification (e.g., for E see Jeyakumar, Carlson, Gunther, & Katzenellenbogen, 2011). That said, the shape of

[2] Stern et al.'s Tables S5 and S6 nicely illustrate our point that simultaneous entry of multiple putative cues is highly insensitive to detecting effects, as each valid cue's effects control for all other valid cues. In Table S5, no *t*-value for moderation of cycle shifts in preferences for 7 cues, entered simultaneously exceeds 0.88 ($p > .381$; mean $p = .61$). Yet, in Table S6, Stern et al.'s composite measure yielded significant moderation. Again, many of the effects in Stern et al.'s multiverse analysis involve simultaneous entry.

[3] We discuss Jünger et al.'s composite measure because it speaks to claims about their own findings. In our view, their composite measure is not a particularly good measure of muscularity within their stimulus set.

[4] Stern et al. state that this finding came from Marcinkowska et al.'s (2018) supplementary analyses. In fact, it was discussed in their text (see p. 117).

the associations between concentrations, binding affinities, and downstream effects on behavior need not follow this pattern (though, physiologically, there is no reason to expect strict linearity). Roney overstates the evidence for linearity. In the example he cites, Bayer, Gläscher, Finsterbusch, Schulte, and Sommer (2018) found a *monotonic* association between E and a hippocampal response. We extracted the data from Bayer et al.'s Fig. 5b (using https://apps.automeris.io/wpd/) and found that strict linear and logarithmic functions fit the association almost identically well. (The correlation between raw and logged E exceeds 0.9.)

To build a data base that assesses the relative predictive power of raw and log-transformed hormone values, researchers may well examine and report associations with both. Roney and Simmons (2013) examined changes in sexual desire as a function of cycle day in 43 women across two cycles, where the E/P ratio identified day of ovulation, a data base suitable to examine day-to-day changes on a psychological variable. We examined correlations of mean sexual desire across days from that study with daily E/P and ln(E/P), using data presented by Roney. Presumably due to lagged effects of E and P (Roney & Simmons, 2013), covariation is maximized when hormone effects are lagged 3 or 4 days. In both cases, correlations for ln(E/P) (0.77, 0.82) exceed those for E/P (0.70, 0.69), contrary to Roney's expectations. (See Table S2, SOM.) But much more data are needed to assess the relative predictive value of raw vs. log-transformed levels.

7. Our target article raised issues concerning scientific strategy, several of which relate to commentaries

Our target article concluded with several observations about scientific strategy. We take this opportunity to briefly expand upon broad observations about strategies that may foster or, conversely, deter scientific progress.

7.1. Embrace uncertainty

Amrhein, Greenland, and McShane (2019) note that it takes a lot of data to estimate true effect sizes and establish boundary conditions. Many times, even “non-significant” effects are not inconsistent with large, theoretically meaningful effects (at the upper bounds of their confidence intervals). Amrhein et al. (2019) encourage researchers to embrace uncertainty, a call for epistemic modesty. Too often, reports reflect “dichotomania”—reported as significant, with confidence intervals, when $p < .05$, and as “non-significant,” with no details, if $p > .05$. Amrhein et al. plead for more “detailed and nuanced” (p. 307) results sections.

We emphasize that our claim about moderation effects in Jünger et al.'s data was modest; our more definitive claim was that Jünger et al. underreported their data and overstated the strength of their conclusions. Even in their response, Stern et al. say they did not report on a preregistered hypothesis because their findings did not lead to “altered conclusions.” They imply that effects were “not significant,” an uninformative binary outcome. In fact, we now know that an analysis using Jünger et al.'s composite measure did yield “significant” moderation. But even if not, sharing this detailed information is important for evaluating the appropriateness of asserting the null hypothesis.

7.2. Preregistration itself does not justify analyses or their meaningfulness

Preregistration justifies and demands that specific analyses be run and reported. But it does not justify their meaningfulness. Too often, in our view, Stern et al.'s justification for particular procedures or interpretation of results lies in the fact that they simply preregistered such procedures or interpretations. For instance, they preregistered 7 features as indicators of muscularity or masculinity. They now cite prior evidence suggesting that these features should be valid indicators and, hence, related to attractiveness and/or formidability. In so doing, they miss the point of our validation analyses: *Attractiveness and formidability do not relate to five of these features in Jünger et al.'s sample of stimuli.* In their own data, these features are not valid for purposes of assessing preferences for cues of upper-body strength—and these features likely do not reflect “good genes.” Appeal to the preregistration does not change that fact.[5]

7.3. “Data-dependent” analyses are sometimes necessary—and their evidentiary value should not be dismissed out of hand because they are data-dependent

Gelman and Loken (2014) explicitly warn of the pitfalls of data-dependent analysis. At the same time, they do not eschew it: “The most valuable statistical analyses often arise only after an iterative process involving the data” (p. 464), illustrated by one of Gelman's own contributions. Earlier, we noted the thoughtful solutions to pitfalls of data-dependent analysis proposed by Jones et al. Their recommendations are valuable. At the same time, they are not the only possible ones. As Gelman and Loken note, awareness of how one's choices can affect results can go a long way toward addressing their impact, as one may then assess whether similar conclusions are reached using other data sources. Independent replication is one obvious possibility. But as an alternative strategy in our target article, we consulted an analysis using Bodily Dominance, a feature not subject to the same selection process as our two-feature composite. The fact that even stronger findings emerged using that measure contradicts the idea that we simply capitalized on chance in constructing the composite. Data-dependent analyses should be evaluated critically, but not dismissed reflexively.

7.4. Even replication studies should be sensitive to empirical patterns that were not expected

From Stern et al.'s commentary, a reader might assume that, in our target article, we dedicated a good deal of space to defending the good genes ovulatory shift hypothesis, as proposed in 1998 and generally represented in the field. Hence, their section 3 (The problem with unfalsifiability), details recent negative evidence. In our own analyses, they note, single women's P has effects opposite to what this hypothesis expects. If we're not willing to accept past and current evidence, they seem to ask, is the hypothesis even falsifiable?

The intense focus on this hypothesis in Stern et al.'s commentary is both puzzling and frustrating, as we do not defend this particular hypothesis. In our section 5.5 (Interpretation), we asked what might explain the pattern suggested. We listed a few potential explanations for the effect within partnered women (*if* it is real), where only one possibility involved good genes. We then observed that the direction of the effect for single women is opposite to what is expected based on the original shift hypothesis. (The dual mating hypothesis may expect no effect in single women, but not an opposite effect.) Hence, we concluded, these findings may “suggest *new* hypotheses about … shifts among single women” (p. XXX; emphasis added). (Parenthetically, we note, it may make sense to expect non-partnered women to be especially cautious about sex during the conceptive phase, but perhaps to be more open to sexual relations that serve functions other than direct conception, such as mate evaluation, when non-conceptive. We note that the idea is post hoc, inspired by the findings, but nonetheless worth exploring.)

[5] In Figure S2, SOM, we present bivariate plots of the 7 features and Bodily Dominance. On upper-to-lower torso ratio, we found one extreme outlier, whose removal enhanced validity of that feature as well as shoulder-to-chest ratio (in a negative direction). Analyses on 3- and 4-feature composites of Strength/Muscularity, with the outlying stimulus figure removed, yielded 3-way interaction effects that further bolster the findings we report for our 2-feature composite and Bodily Dominance. See Table S3.

The structured, narrow aims of preregistered replication work may inhibit attention to novel findings and interpretation. While these aims are needed, ultimately a study should speak to empirical phenomena that exist, whether expected by existing theory or not. Precisely because they cannot be explained by extant theory, unexpected findings inspire theoretical development.

It is reasonable and necessary to critically evaluate novel findings. For instance, one should desire to see replication. Rather than dismissal, unexpected findings may warrant very cautious, critical entertainment. One way that unexpected findings can be critically entertained, even prior to replication, is in light of other findings or theories in the field. Recent, large replication studies have found little support for predictions from the ovulatory shift hypothesis (Stern et al.). At the same time, multiple studies now detect relationship status moderation of P associations with preferences (target article, Section 5.7, footnote 18). Effects for single women are on balance as strong as effects (in the opposite direction) for partnered women (see, for instance, Marcinkowska et al., 2018, Fig. 2). Currently, no explanation for these effects has been offered. Though some associations have been found with within-cycle variation in P, other studies find associations with variation in P across women. They could have nothing to do with one another. At the same time, Stern et al. note that power to detect moderation by relationship status in these studies is likely modest, which means that real effects will often not be detected ("significant"). Will interesting patterns of moderation by relationship status turn out to be systematic, and importantly inform theories about psychological shifts across the cycle and their functional significance? Perhaps yes, perhaps no. More empirical work is needed to explore them, and more theoretical work is needed to explain them, should they be real. Premature dismissal of unexpected findings discourages that work.

Roney wonders "what we should consider an 'effect' at all," and suggests that the tests of special interest "we should be focused on depend on how specific effects relate to specific theoretical positions." We agree that, most importantly, effects inform and constrain theory. But that is precisely the reason one should attend to empirical patterns, whether expected by existing theory or not. Not only may they lead to new ways of thinking; they speak to existing theories. For instance, Roney and Simmons' motivational priorities theory does not predict the 3-way interaction we report. If it does turn out to be robust, then, this effect suggests that the theory is not a complete explanation of cycle shifts in sexual interests.

Naturally, theories must be falsifiable. The "ovulatory shift hypothesis" is not a catch-all theory that can explain any hormonally mediated shift in sexual interests. The actual ovulatory shifts that exist naturally constrain the content of appropriate explanation. Our appeal for *new* theory implies that existing theory may be in need of revision.

8. Conclusion: just say no to just saying no

Jünger et al. (2018) proclaimed a null hypothesis: their data left "no room" for cycle shifts in mate preferences for masculine characteristics. Though our target article concerned a particular moderation effect in their data, one preregistered but not assessed in their paper, we highlighted broader themes. First, researchers should not *say no*, there exist no meaningful effects, without evidence that goes well beyond simple hypothesis testing (e.g., equivalence testing). Amrhein et al. (2019) emphasize this point. Second, when one reports a null effect, one should not *just* say no, we found no effect. As also emphasized by Amrhein et al. (2019), detailed analyses are needed, and that holds for both "significant" and "non-significant" effects. Stern et al. admit to holding back on reporting critical analyses that would have allowed their readers to better evaluate their bold null claims. Third, when one observes patterns that were not expected, long-run scientific progress does not benefit when researchers just dismiss those effects because they were not expected. In many individual instances, of course, it may well be that unexpected patterns are unreliable. But the ones that are real, even if the small minority, may importantly shape theoretical understanding.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.evolhumbehav.2019.08.008.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Comment: Retire statistical significance. *Nature, 567*, 305–307.

Bayer, J., Gläscher, J., Finsterbusch, J., Schulte, L. H., & Sommer, T. (2018). Linear and inverted U-shaped dose-response functions describe estrogen effects on hippocampal activity in young women. *Nature Communications, 9*, 1–12. https://doi.org/10.1038/s41467-018-03679-x.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*, 460–465.

Jeyakumar, M., Carlson, K. E., Gunther, J. R., & Katzenellenbogen, J. A. (2011). Exploration of dimensions of estrogen potency. *Journal of Biological Chemistry, 286*, 12971–12982.

Jones, B. C., Hahn, A. C., Fisher, C., Wang, H., Kandrik, M., & DeBruine, L. M. (2018). General sexual desire, but not desire for uncommitted sexual relationships, tracks changes in women's hormonal status. *Psychoneuroendocrinology, 88*, 153–157.

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*, 412–423. https://doi.org/10.1016/j.evolhumbehav.2018.03.007.

Marcinkowska, U. M., Kaminski, G., Little, A. C., & Jasienska, G. (2018). Average ovarian hormone levels, rather than daily values and their fluctuations, are related to facial preferences among women. *Hormones and Behavior, 102*, 114–119.

Roney, J. R., & Simmons, Z. L. (2013). Hormonal predictors of sexual motivation in natural menstrual cycles. *Hormones and Behavior, 63*, 636–645.

Sollberger, S., & Ehlert, U. (2016). How to use and interpret hormone ratios. *Psychoneuroendocrinology, 63*, 285–297.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*, 702–712.

West, S. G., Ryu, E., Kwak, O.-M., & Chan, H. (2011). Multilevel modeling: Current and future applications in personality research. *Journal of Personality, 79*, 1–50.