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

Published version:

DOI:10.1016/J.EVOLHUMBEHAV.2019.08.008

(Article begins on next page)

Psychological Cycle Shifts Redux: Revisiting a Preregistered Study Examining Preferences for Muscularity

Steven W. Gangestad[1], Tran Dinh[1], Nicholas M. Grebe[2],

Marco Del Giudice[1], & Melissa Emery Thompson[3]

[1]Department of Psychology, University of New Mexico

[2]Department of Evolutionary Anthropology, Duke University

[3]Department of Anthropology, University of New Mexico

Send correspondence to Steve Gangestad, Department of Psychology, University of New Mexico, Albuquerque, NM 87131 email: sgangest@unm.edu

RUNNING HEAD: Hormone-Associated Shifts Redux

14 May 2019

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., Roney & Simmons, 2013; Arslan et al., in press; Jones et al., 2018a)—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 et al., 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 et al., 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 et al., 2014a); the other detected no such effects (Wood et al., 2014; cf. Gildersleeve et al., 2014b).

Based on additional meta-analytic analyses, Gangestad et al. (2018a) 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., 2018). Still, they emphasize, more research is needed. Among promising candidates for cycle shifts are preferences for muscular features. Jünger et al. (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 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. XXX; 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 et al., 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 et al., 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.

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

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

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]

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.

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

[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.

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 et al., 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).

---

[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 .15 to .35; a near-identical drop in power occurred.

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.

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 et al., 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.

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. (2018b) 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 criticisms of their analyses.[5]

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

[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.

(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 et al., 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

[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 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.)

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 et al., 1995). We proposed to run both sets of analyses.

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 et al., 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.

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,

[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 = -.60, -.73 for raw and logged progesterone values, respectively, and .38, .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.

compromising 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 et al. (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 composite does not fully capture

[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.

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 et al., 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]

---

[10] This composite correlates .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 .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

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.

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

---

relatively weak, $r = .20$, $p = .073$. 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.

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.

[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.

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 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

[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.

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

[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.

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 Tables 5 and 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-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 5^{th} and 95^{th} 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, 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 Figure 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 Tables 7 and S14 and Figure 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 (α = .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 Figure 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 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 et al., 2018b), 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 and Hoogland [1982] found that the initial LH surge, measured in serum, occurred 11-24 hours 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 et al., 2001) (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. XXX); 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. XXX; 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 et al., 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 et al., 2012; Barr et al., 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. XXX; 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

[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.

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 (-1.32s and 1.55s 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 et al., 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., Roney & Simmons, 2013; Grebe et al., 2016).

5.7. An independent demonstration

Since we conducted these analyses, we learned of another, recently published study that found a similar interaction. Marcinkowska et al. (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]

[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.

[18] Both Marcinkowska et al. (2018) and DeBruine, Hahn, and Jones (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

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

---

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, Greenland, & McShane, 2019). The issue of whether these interactions are related and due to direct effects of progesterone is, in our view, not yet fully resolved.

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.

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 Amrheim et al., 2019).

Preregistration and up-down thinking in hypothesis-testing. As argued by others (e.g., Cumming, 2014; Amrhein et al., 2019), 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.

[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.

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, 1934] 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).

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 et al., 2017). Preregistered analyses address targeted 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.

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 et al., 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 et al., 2012; Fiedler, 2017). 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.

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. In *American Journal of Physical Anthropology* (Vol. 159, pp. 78-78). Hoboken NJ: Wiley-Blackwell.

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: 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.

Bates, D., Kleigl, R., Vashisth, S., & Baayen, H. (2015) Parsimonious linear models. Available from arXiv:1506.04967 (stat.ME).

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.

Dunson, D. B., Weinberg, C. R., Baird, D. D., Kesner, J. S., & Wilcox, A. J. (1999). Assessing human fertility assessing several markers of ovulation. *Statistics in Medicine*, *20*, 965-978.

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.

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

Gangestad, S. W., Dinh, T., Grebe, N. M., Gildersleeve, K., Emery Thompson, M., & Haselton, M. G. (2018a). 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

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

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.

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*.

Jones, B. C., Hahn A. C., Fisher, C. Wang, H. Kandrik, M. Han C., Fasolt, V., Morrison, D. K. Lee, A., Holzleitner, I. J. Roberts, S. C., Little, A. C., & DeBruine, L. M. (2018b). Women's preferences for facial masculinity are not related to their hormonal status. *Psychological Science*.

Jones, K. A. (1996). Summation of basic endocrine data. In Gass, G. A., Kaplan, H. M. (Eds.), *Handbook of Endocrinology, Volume 1*, second edition, 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. doi: 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*.

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 edition. New York: McGraw-Hill/Irwin.

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

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.

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

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.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., Bates, D. (2017). Balancing Type I error and power in linear models. *Journal of Memory and Language*, *94*, 305-315.

Maxwell SE, Lau MY, Howard GS. 2015. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist,* 70, 487–98

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*. DOI: 10.11.1126/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.

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

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, *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. doi: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., Feinberg, D. R., Sharp, M. A., & Al-Dujaili, E. A. (2007). Raised salivary testosterone in women is associated with increased attraction to masculine faces. *Hormones and Behavior*, *52*, 156-161.

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.

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.

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.

Table 1

Jünger et al.'s Data: Sexual Attractiveness and Bodily Dominance in Relation to Male Features

	Predicting Sexual Attractiveness			Associations with Bodily Dominance	
	γ / SE	t	p	*r*	*r* w BMI controlled
BMI	-.78/.2	-3.79	**<.001**		
Strength	.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 $\underline{\gamma} \pm 2 \times$ 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.

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 (95^{th} and 5^{th} 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).

Table 4

Results of Multilevel Regression Analyses on Jünger et al.'s Data: Predictors of Sexual Attractiveness

	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*
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 x Relationship Status	-0.22/0.08	-2.71	**0.008**				-.37/.10	-3.57	**<.001**			
mean E/P x Relationship Stat	.09/.10	0.82		-.11/.07	-1.67	*0.094*						
mean T x Relationship Status	-0.11/0.11	-0.93										
BMI x Relationship Status	-0.09/0.06	-1.64	0.101	-.08/.06	-1.43	0.153	-.03/.05	-0.55				
BMI x ww E/P	-0.01/0.01	-0.54		-.00/.01	-0.44		-.01/.01	-0.65				
BMI x ww T	.02/.01	1.58	0.114				.03/.02	2.10	**0.036**			
BMI x mean E/P	-0.02/04	-0.58		-.02/.04	-0.47							
BMI x mean T	.06/.04	1.50	0.135									
S/M x Relationship Status	.03/.05	0.70		.03/.05	0.57		.03/.04	0.61		.02/.03	0.55	
S/M x ww E/P	-.00/.01	-0.29		-.00/.01	-0.34		-.00/.01	-0.15		-.00/.01	-0.34	
S/M x ww T	-0.01/0.01	-1.09					-.02/.02	-1.52	0.129			
S/M x mean E/P	.01/.03	0.28		.00/.02	0.13							
S/M x mean T	-0.03/0.03	-1.13										

Rel Stat x BMI x ww E/P	-.02/.02	-1.22	0.222	-.02/.02	-1.09		-.04/.02	-1.78	*0.074*			
Rel Stat x BMI x ww T	.03/.02	1.45	0.146				.02/.03	0.56				
Rel Stat x BMI x mean E/P	-.16/.05	-3.26	**0.001**	-.16/.05	-3.24	**0.001**						
Rel Stat x BMI x mean T	-0.05/0.05	-1.04										
Rel Stat x S/M x 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 x S/M x ww T	-0.02/0.02	-1.16	0.246				-.00/.03	-0.12				
Rel Stat x S/M x mean E/P	.06/.04	1.34	0.179	.06/.04	1.42	0.155						
Rel Stat x 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 $\underline{\gamma} \pm 2 \times SE$.

[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		
	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 x Relationship Status	-.13/.12	-1.08		-.03/.12	-1.08	
P x Relationship Status	.04/.05	0.68		.04/.05	0.69	
BMI x Relationship Status	-.03/.05	-0.52				
BMI x E	-.02/.01	1.41	0.159			
BMI x P	.04/.06	0.91				
S/M x Relationship Status	.03/.04	0.63		.00/.05	0	
S/M x E	-.02/.01	-1.58	0.114	-.01/.01	-1.46	0.145
S/M x P	-.00/.01	-0.25		-.00/.01	-0.19	
Rel Stat x BMI x E	-.03/.03	1.2	0.229			
Rel Stat x BMI x P	.05/.02	2.29	**0.022**			
Rel Stat x S/M x E	.01/.03	0.38		.00/.02	0.23	
Rel Stat x 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 x E/P	.01/.02	0.79		-.03/.02	-1.74	0.083						
BMI x T	.02/.02	1.1		.04/.02	1.97	**0.049**						
S/M x E/P	-.03/.02	-2.05	**0.041**	.03/.02	1.87	*0.061*						
S/M x 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 x E	.00/.02	0.16		.03/.02	1.78	*0.075*						
BMI x P	-.02/.02	-0.95		.04/.02	2.31	**0.021**						
S/M x E	-.02/.02	-1.45	0.148	-.02/.02	-0.83							
S/M x 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 Tables 3 and 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 $\underline{\gamma} \pm 2 \times \mathrm{SE}$.

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 x E/P	-.03/.07	-0.42		-.02/.07	-0.37	
Relationship Status x T	-.38/.10	-3.59	**<.001**	-.37/.10	-3.58	**<.001**
BMI x Relationship Status	- .04/.05	-0.92		-.06/.06	-1.02	
BMI x E/P	-.00/.01	-0.02		.01/.01	0.09	
BMI x T	.02/.01	1.40	0.162	.03/.02	1.78	*0.075*
BD/SF x Relationship Status	.09/.05	1.73		.06/.05	1.14	
BD/SF x E/P	-.02/.01	-2.36	**0.018**	-.01/.01	-1.29	0.196
BD/SF x T	-.00/.01	-0.07		-.02/.02	-1.01	
Rel Stat x BMI x E/P	-.02/.02	-1.17	0.24	-.05/.02	-0.19	
Rel Stat x BMI x T	.01/.03	0.55		.02/.03	0.63	
Rel Stat x BD/SF x 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 x E	-.13/.11	-1.08		-.13/.12	-1.08	
Relationship Status x P	.04/.06	0.75		.04/.06	0.68	
BMI x E	.02/.01	1.45	0.147	.02/.01	1.81	*0.071*
BMI x P	.00/.01	0.32		.00/.01	0.31	
BD/SF x E	-.03/.01	-2.68	**0.007**	-.03/.01	-2.29	**0.022**
BD/SF x P	.01/.01	1.52	0.130	.01/.01	0.63	
Rel Stat x BMI x E	.03/.02	1.54	0.123	.03/.03	1.09	
Rel Stat x BMI x P	.03/.02	1.78	*0.074*	.06/.02	2.72	**0.007**
Rel Stat x BD/SF x E	.01/.02	0.5		.01/.03	0.52	
Rel Stat x BD/SF x 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 $\underline{\gamma}$ ± 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.

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 Tables 6, 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 Tables 6, S15)												

	.07/.02	3.39	**0.001**	.06/.02	3.24	**0.001**	.08/.02	3.54	**<.001**	.05/.02	3.41	**0.001**
<u>**Hormonal predictors: estradiol and progesterone entered separately**</u>												
Ln(E) and ln(P) entered as hormonal predictors (from Tables 4, 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 $\underline{\gamma} \pm 2 \times$ 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

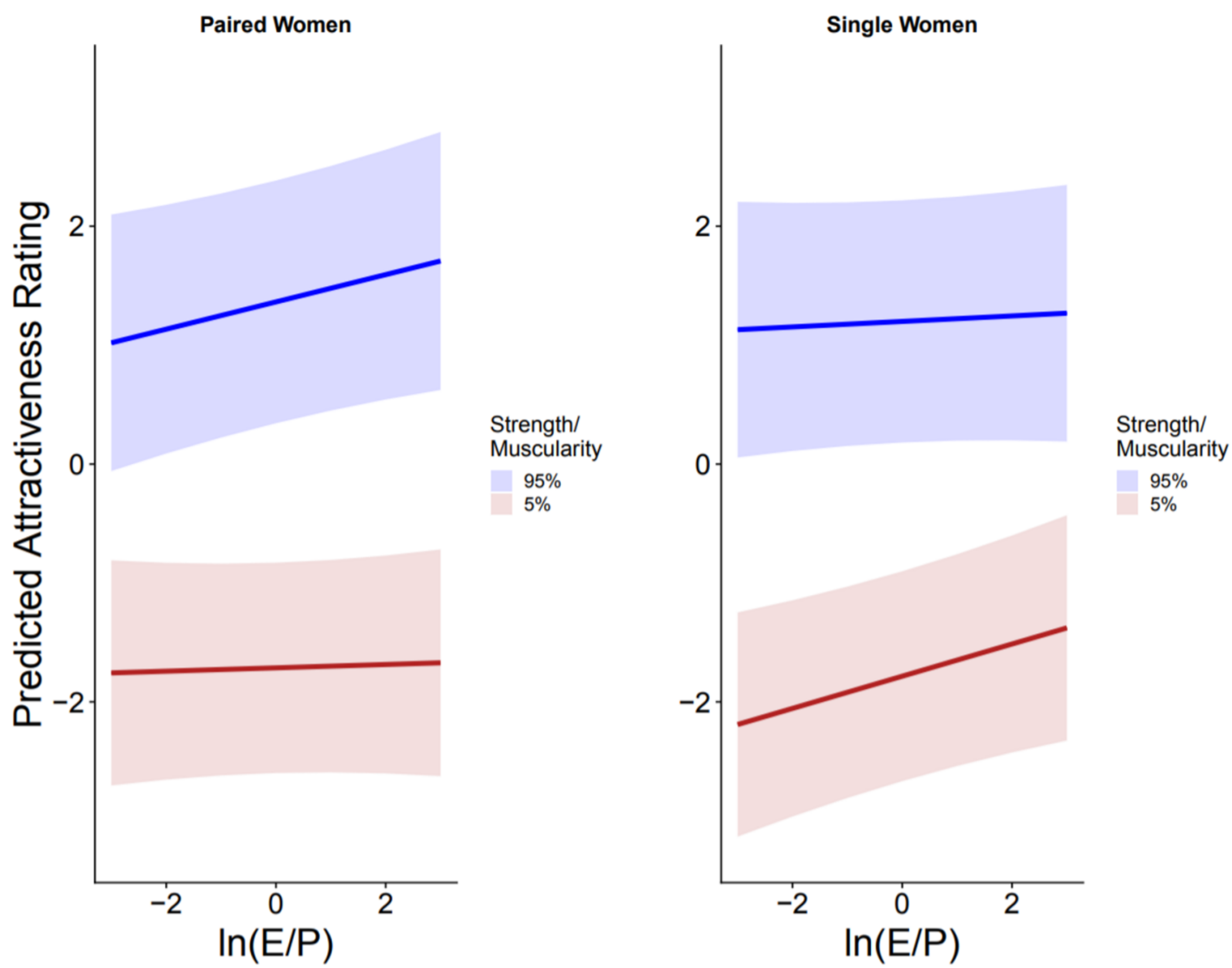
	γ / SE	*t*	*p*
BMI	-1.10/.25	-4.39	**<.001**
Strength/Muscularity (S/M)	1.00/.29	3.49	**<.001**
Relationship Status	.20/06	-3.54	**<.001**
Cycle Phase	.07/.04	2.09	**0.037**
Phase x Relationship Status	.12/.06	1.95	*0.051*
BMI x Relationship Status	-.03/.05	-0.61	
BMI x Phase	-.02/.02	-0.28	
S/M x Relationship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

Notes. All quantitative predictors z-scored. Relationship status effect coded: single = -.5, partnered = .5. Phase effect codes: -.5 = luteal; .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 $\gamma \pm 2 \times$ SE.

Figure Caption

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

Figure 1.


Paired Women
Single Women
Predicted Attractiveness Rating
2
0
−2
−2
0
2
ln(E/P)
Strength/
Muscularity
95%
5%