Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.elsevier.com

Commentary

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

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

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

ARTICLE INFO

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

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

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

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

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

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

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

There are many ways to generate, from a decision tree, a collection of weak, unjustifiable tests. An invalid measure may substitute for a valid one. A valid measure can be split into unreliable components. Or,

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

https://doi.org/10.1016/j.evolhumbehav.2019.08.008
Available online xxx

one can enter multiple correlated valid indicators of the same trait simultaneously, which fractionates valid variance captured by each. Consider an example. Suppose the predictor of a criterion is a personality trait. This trait has been assessed with 10 items, but only 5 turn out to be valid. The best measure of the trait is a composite of those 5. But of course, one can generate many alternative models: e.g., using the total sum of 10 entered; entering items separately; entering items simultaneously. In the latter analysis, even each valid item likely has very little residual validity, as its valid variance overlaps with that of the remaining 4, which is partialled out. The combinatorial nature of decisions means that many weakly powered tests can be generated from few decisions. It would be nonsensical to argue that an effect found with the optimal composite is not robust because its effects are dwarfed by those of very weak alternatives.

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

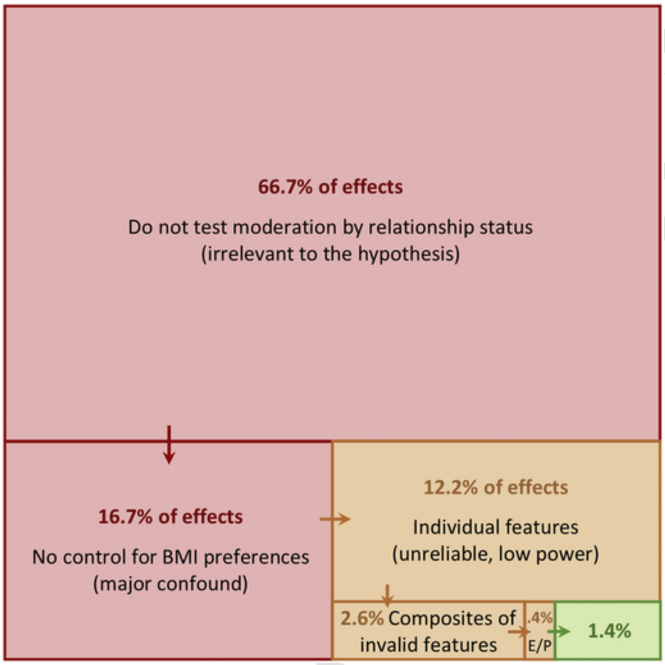

Fig. 1. The distribution of effects in Stern et al.'s multiverse. Starting at the top, 66.7% of effects concern a hypothesized effect separate from the moderation effect; separate conceptual effects demand separate multiverse analyses. Half of the 33.3% that remain 16.7% of the total do not control for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

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

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

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

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

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

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

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

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

Stern et al. paint our findings as highly selective and thus not robust. However, they ignore a critical set of our analyses. We evaluated the validity of cues of male muscularity by examining their association with independent observers' ratings of Bodily Dominance (formidability), which correlate highly with bodily attractiveness. Though not a perfect criterion of muscularity, Bodily Dominance (with BMI controlled) is likely a better measure than any single physical cue, and likely a better measure than our composite of two cues. From viewing 3D bodies, raters could perceive more than just a few features when judging Bodily Dominance. It follows that analyses should show similar or stronger results when substituting our muscularity composite with Bodily Dominance and they did. In parallel multiverse analyses (Fig. 2), all p for hormonal analyses were 0.008; p for cycle phase were 0.072 (full sample) and 0.048 (preregistered sample of 112).

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

2.2. We performed additional analyses for robustness checks

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

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

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

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

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

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

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

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

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

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

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

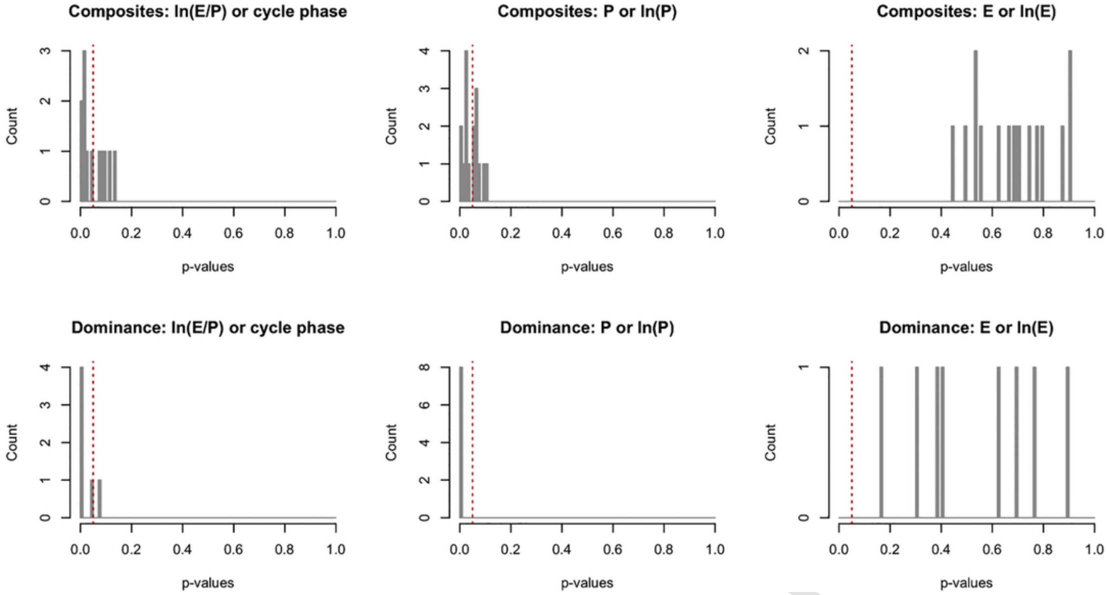

Fig. 2. Distribution of *p*-values in a multiverse of effects using body feature composites (top panels) and Bodily Dominance ratings (bottom panels). Two composites are used: our empirically vetted composite of Strength/Muscularity; and a factor tapping this dimension extracted from all 7 male features.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

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

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

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

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

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

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

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

We offer a few reflections on Roney's claims about raw vs. log-transformed hormone ratios. (a) Though the E/P ratio may track conception risk very well, that need not imply that the physiological and behavioral effects of E and P are captured by the ratio. The ratio peaks near maximal conception risk a day or two prior to ovulation because of its temporal association with the event of ovulation itself, not because peak E/P exerts immediate causal effects on it (or, for that matter, adaptive behavior). As Higham notes and we discuss above, the effects of E and P need not perfectly track conception risk. (b) The fact that the E/P ratio reflects complex non-linear interactions between E and P offers no assurance that the ratio *appropriately* captures true *E*-P interaction effects. As Sollberger and Ehlert (2016) advise, researchers should model interaction effects (e.g., by inclusion of E × P terms) rather than blindly assume that a hormone ratio captures interaction effects. (c) A widely-adapted model argues that ligands' binding affinities are a sigmoidal function of the log of the availability of the ligand (e.g., hormone concentration), with some justification (e.g., for E see Jeyakumar, Carlson, Gunther, & Katzenellenbogen, 2011). That said, the shape of the associations between concentrations, binding affinities, and downstream effects on behavior need not follow this pattern (though, physiologically, there is no reason to expect strict linearity). Roney overstates the evidence for linearity. In the example he cites, Bayer, Gläscher, Finsterbusch, Schulte, and Sommer (2018) found a *monotonic* association between E and a hippocampal response. We extracted the data from Bayer et al.'s Fig. 5b (using https://apps.automeris.io/wpd/) and found that strict linear and logarithmic functions fit the association almost identically well. (The correlation between raw and logged E exceeds 0.9.)

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

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

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

7.1. Embrace uncertainty

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

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

7.2. Preregistration itself does not justify analyses or their meaningfulness

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

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

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

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

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

The intense focus on this hypothesis in Stern et al.'s commentary is both puzzling and frustrating, as we do not defend this particular hypothesis. In our section 5.5 (Interpretation), we asked what might explain the pattern suggested. We listed a few potential explanations for the effect within partnered women (*if* it is real), where only one possi-

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

bility involved good genes. We then observed that the direction of the effect for single women is opposite to what is expected based on the original shift hypothesis. (The dual mating hypothesis may expect no effect in single women, but not an opposite effect.) Hence, we concluded, these findings may suggest *new* hypotheses about shifts among single women (p. XXX; emphasis added). (Parenthetically, we note, it may make sense to expect non-partnered women to be especially cautious about sex during the conceptive phase, but perhaps to be more open to sexual relations that serve functions other than direct conception, such as mate evaluation, when non-conceptive. We note that the idea is post hoc, inspired by the findings, but nonetheless worth exploring.)

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

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

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

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

8. Conclusion: just say no to just saying no

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

Appendix A. Supplementary data

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

References

Amrhein, V., Greenland, S., McShane, B., 2019. Comment: Retire statistical significance. Nature 567, 305 307.

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

Gelman, A., Loken, E., 2014. The statistical crisis in science. American Scientist 102, 460 465.

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

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

Jünger, J., Kordsmeyer, T.L., Gerlach, T.M., Penke, L., 2018. Fertile women evaluate male bodies as more attractive, regardless of masculinity. Evolution and Human Behavior 39, 412 423. https://doi.org/10.1016/j.evolhumbehav.2018.03.007.

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

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

Sollberger, S., Ehlert, U., 2016. How to use and interpret hormone ratios. Psychoneuroendocrinology 63, 285 297.

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

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