The evidence for good genes ovulatory shifts in Arslan et al. (2018) is mixed and uncertain

Ruben C. Arslan

Personality Psychology and Personality Diagnostics, University of Leipzig, Leipzig, Germany

Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

Julie C. Driebe

Biological Personality Psychology, University of Goettingen, Germany

Julia Stern, Tanja M. Gerlach and Lars Penke

Biological Personality Psychology, University of Goettingen, Germany

Leibniz ScienceCampus Primate Cognition, Goettingen, Germany

Author Note: The authors thank Linda Kerbl for statistical proofreading, and Malte Elson, Julia M. Rohrer, Anne M. Scheel, and Sarah J. Lennartz for helpful comments on an earlier version of this draft.

Supporting materials can be found on OSF at https://osf.io/t57nv/

This article was published in JPSP with the DOI http://dx.doi.org/10.1037/pspp0000390.

Correspondence concerning this article should be addressed to Ruben C. Arslan,

E-mail: ruben.arslan@gmail.com

Abstract

In Arslan et al. (2018), we reported ovulatory increases in extra-pair sexual desire, in-pair sexual desire, and self-perceived desirability, as well as several moderator analyses related to the good genes ovulatory shift hypothesis, which predicts attenuated ovulatory increases in extra-pair desire for women with attractive partners. Gangestad and Dinh (2021) identified errors in how we aggregated two out of four main moderator variables. We are grateful that their scrutiny uncovered these errors. After corrections, our moderation results are more mixed than we previously reported and depend on the moderator specification. However, we disagree that the evidence for moderation is robust and compelling, as Gangestad and Dinh (2021) claim. Our data are consistent with some previously reported effect sizes, but also with negligible moderator effects. We also show that what Gangestad and Dinh (2021) call an "a priori [...] more comprehensive and valid composite" is poorly justifiable on a priori grounds, and follow-up analyses they report are not robust to a composite specification that we consider at least as reasonable. Psychologists have to become acquainted with techniques such as cross-validation or training and test sets to manage the risks of data-dependent analyses. In doing so, we might learn that we need new data more often than we intuit and should remain uncertain far more often.

Keywords: ovulatory cycle shifts, robustness, data-dependent analyses

Transparency can only improve science insofar it enables scrutiny. We are glad that Gangestad and Dinh's (2021) scrutiny of our open code and shared data led to the correction of two important errors in our data cleaning process, which affected two of the four main moderator variables we examined.

We agree with Gangestad and Dinh (2021) on many points. We agree that our data shows ovulatory cycle increases, exclusively among women who did not use hormonal contraception, in extra- and in-pair desire, and self-perceived desirability. We agree that our preregistration was underconstrained and contained suboptimal procedures. We agree that our mate retention measure was poor (but see Note S1). We agree that our fertile window estimator is less valid than estimators based on luteinising hormone tests and that it is unclear if our much larger sample size could compensate for this flaw. We agree that not modelling random slopes for the fertile window effect introduces a liberal bias (Barr et al., 2013). Most importantly, we agree that binary summaries of evidence are inappropriate, that our design was not ideally suited to falsify the good genes ovulatory shift hypothesis, and that we could not conclusively rule out non-negligible moderation. In brief, "mixed and uncertain evidence" would have better captured the uncertainty in our moderation results than the summary "no evidence" in Arslan et al. (2018), which Gangestad and Dinh (2021) took issue with.

Coming to our disagreements, we see mixed and uncertain evidence where Gangestad and Dinh (2021) see "robust evidence for moderation of ovulatory shifts" in extra-pair desire and partner mate retention. To elucidate this disagreement, we have to clarify interpretive issues related to preregistration and *post-hoc* analyses and revisit[1] the results after both coding errors identified by Gangestad and Dinh (2021) have been corrected (the errors were jumbling values for relative partner attractiveness and treating missing values in satisfaction with sex for women who had not yet had sex with their partners as zero satisfaction).

An underconstrained preregistration

Preregistrations mainly serve as transparent constraints on data-dependent analyses. Our effort to constrain ourselves was challenged in several ways. First, research published after our preregistration suggested improvements on our preregistered procedure. Second, we planned to replicate several related studies and operationalizations at once, but did not plan for the likely inflation of the false positive rate (see Table S3). Third, our preregistration did not anticipate all decisions to be made, especially with respect to data cleaning, and left other decisions, such as exclusion criteria, vague and open-ended.

To address the first challenge, we amended our preregistration during data collection but before analysis. We decided to examine additional exclusion criteria,

[1] We learned about the first error via personal correspondence, but only learned about the second error through the publication of Gangestad and Dinh (2021), so we do not stand by our previous corrected summary (Arslan et al., 2019).

an alternative, broader fertile window predictor, as well as to add plots using more data. During analysis, we became convinced that continuous conception risks improved on windowed predictors, but only used continuous predictors in non-confirmatory robustness analyses. However, we decided against a post-hoc modification of a central outcome, mate retention (see Note S3).

To address the second challenge, we decided on a lower significance threshold (.01 rather than .05) for preregistered analyses and showed that interpretations of the main effects did not hinge on the chosen threshold or multiple testing correction. For moderator tests, as our uncorrected preregistered tests (Arslan et al., 2018) were insignificant, we did not discuss multiple testing. To trim the length of the manuscript to a readable size, we reported many details, including most information on moderator analyses, in an online supplement. Regrettably, we then summarized the evidence on moderation inadequately.

We did not appropriately address the third challenge. We should have listed under- constrained decisions, noted how and when we decided, and explicitly checked robustness to other justifiable decisions. We conducted extensive robustness analyses, but failed to transparently link the analyses to vagueness in our preregistration. We also failed to note which decisions were made blind to the final results (e.g., most decisions on data cleaning, operationalization, and exclusion criteria) and which decisions we made later (e.g., certain *post-hoc* robustness analyses). For instance, to determine whom to exclude from the dataset, we explicitly preregistered that we would *examine* the effect of different exclusion criteria, but not how. In Arslan et al. (2018), we adopted a single set of criteria for the preregistered analyses and reported other possible sets as robustness analyses (see Table S3). Further, we neither preregistered how we would operationalize moderator constructs, nor how we would analyze them, nor how we would interpret inconsistent results. Previous literature had rarely repeated the exact same items for testing ovulatory shifts (Harris et al., 2013); this lack of direct replication motivated our study. We made an effort to replicate previously used items closely, but in one case (partner short-term relative to long-term attractiveness) we chose an operationalization that we considered a better implementation of the theoretical prediction over one that more directly followed Haselton and Gangestad (2006), as explained in Note S3. Our preregistration should have made this explicit, but did not.

A corrected summary of the corrected evidence for moderation

To test the good genes ovulatory shift hypothesis (GGOSH), our preregistered analyses specified four moderators (partner physical attractiveness, partner short-term attractiveness, partner's relative attractiveness compared to self, partner short-term relative to long-term attractiveness) of ovulatory change (operationalized as the effects of our narrow fertile window predictor) on three outcomes (extra-pair desire and behaviour, in-pair desire, male mate retention). In addition, we tested three different operationalizations of partner's short-term relative to long-term attractiveness (a difference score, dual moderators, and a three-way interaction, see

Note S3) for a total of 6 different moderator tests per outcome and 18 in total. In our corrected preregistered analyses, none of the 18 moderator tests were statistically significant in the predicted direction at the $p < .05$ level. As shown in Table S1, although four moderator tests were significant at $p < .05$ for in-pair desire, all effects were in the opposite direction of prediction – that is, women who reported higher partner attractiveness reported smaller ovulatory increases in in-pair desire. All moderators for extra-pair desire and behavior (EPDB) and for partner mate retention changes were in the predicted direction – that is, women who reported higher partner attractiveness reported smaller ovulatory increases in EPDB and partner mate retention, but none were significant (see also Note S1).

We improved on the preregistered procedures in several ways, chiefly by including more women and more days per woman and a more appropriate model in additional robustness analyses (see Table S2 and Note S2). Here, results were more mixed than in the preregistered analyses. For in-pair desire, moderators were close to zero and inconsistent in direction. For mate retention, moderators were also close to zero and mixed in direction; for the relative attractiveness specification, the effect was in the predicted direction, and the 95% but not the 99% CI excluded zero. For extra-pair desire and behaviour, all moderator estimates were descriptively in the predicted direction. Effect sizes varied twofold depending on the moderator specification. For the short-term attractiveness specifications (but not physical and relative attractiveness), several 99% and 95% confidence intervals excluded zero.

On balance, across all outcomes and all corrected analyses (preregistered and robustness) the confidence intervals include some previously reported GGOSH effects, but also negligible effect sizes, and, depending on the moderator specification, effects counter to GGOSH predictions. As we understand it, the evidence for GGOSH moderation is not robust (i.e., insensitive to minor differences in measures and analyses) but mixed and uncertain, and consequently requires further investigation in new data.

Arbitrary or "more comprehensive and valid"?

Going beyond correcting our analyses, Gangestad and Dinh (2021) introduce a 5-component composite moderator, which they describe as follows: "on a priori basis [...] a more comprehensive and valid composite measure of male partner sexual attractiveness" (p. 12). Gangestad and Dinh (2021) report that they formed this measure to estimate the joint probability of all effect sizes under the null hypothesis. In aggregating this composite, they deviated from our aggregation procedure. We think it is important for post-hoc deviations to be transparent, justifiable and mutually consistent. To this end, we list all partner attractiveness items in Table 1 and discuss all deviations by Gangestad and Dinh (2021) from our decisions in the following.

First, we address decisions about individual items. **a)** In their short-term/sexual attractiveness composite, Gangestad and Dinh (2021) include an item (ST1) from our mate value scale that is clearly about *long-term* attractiveness ("How

difficult would it be for your partner to find another partner for a long-term relationship, who is as desirable as you?"). Further, this item and a parallel item about short-term attractiveness (ST2) were poorly constructed: They involved counterfactuals, a comparison to own mate value, and had low correlations with other items in the scale (see Table 1). We had noted these problems in our code comments and showed results both with and without these problematic items (Arslan et al., 2018, SOM). **b)** In our relative attractiveness scale, we had aggregated an item about partner attractiveness relative to own attractiveness (ST6, Table 1) with a difference score of partner and own mate value. Gangestad and Dinh (2021) include the former, relative item, but also include *non*-relativised partner mate value (ST3-ST5). They then use their composite in an analysis where the partner's long-term attractiveness is adjusted for. So, they mix two different types of relativized measures in their analysis (partner attractiveness relative to own attractiveness and short-term/sexual attractiveness relative to long-term attractiveness). We find this confusing, difficult to interpret, and note the comparatively low item-scale correlation for item ST6 (Table 1). **c)** Gangestad and Dinh (2021) include the sexual satisfaction item (ST7), which has missing data. In one specification, they impute average sexual satisfaction to women who never had sex with their partner. In a second specification, they omit these women. They do not consider a third specification, omitting the item from the composite, as we had done in our robustness analyses (Arslan et al., 2018, SOM). We chose the third specification, because of the low correlation of the sexual satisfaction item with sexual attractiveness items, the missing data, and its divergent content. **d)** They exclude an attractiveness item that we presented as part of our robustness analyses in Arslan et al. (2018, SOM, "How attractive is your partner for other women, compared to other men?", ST8), even though it is very similar to the partner mate value items in content and correlates highly with the other scale items.

Second, we address how and when items were aggregated: **e)** Whereas we weighted items equally (i.e., by the item variance) and did not standardise components except in one case where response scales differed, Gangestad and Dinh (2021) standardise all components, even those including different numbers of items, which implies, for instance, that the physical attractiveness items are downweighted in their composite compared to the attractiveness relative to self item. **f)** Finally, Gangestad and Dinh (2021) formed a "more comprehensive" composite for short-, but not for long-term attractiveness. In previous, comparable studies, Haselton and Gangestad (2006) and Pillsworth and Haselton (2006) both averaged a rating item of long-term attractiveness and items about social status, current and future financial prospects. We had planned to do the same, but internal consistency analysis and factor analysis for all long-term attractiveness items (a rating item, financial status, occupational prospects, net income, LT1-LT4, Arslan et al., 2018, SOM) showed that the rating item (LT1) was not highly correlated with the other items (LT2-LT3), causing the scale's Cronbach's alpha to fall below our preregistered criterion of .60. Therefore, we opted not to use the aggregate. Instead, we used the

rating item (LT1) in our main analyses and reported an aggregate of LT2-LT4 as a moderator in our robustness analyses to make all patterns transparent. Given Gangestad and Dinh's (2021) professed intention to estimate the joint probability under the null, we find it inconsistent to form a composite of heterogeneous items for short-term, but not long-term attractiveness.

For some of these deviations (e.g., **c**), we do not claim that our original approach was superior, although Gangestad and Dinh's (2021) deviations from our approach strike us as arbitrary. Others (e.g., **b)**, **d)**, **e), f)**) we find mutually inconsistent and **a)** simply seems wrong. We do not find that Gangestad and Dinh (2021) support their claims about their composite. By what measure is their composite "more [...] valid" than what we had reported? It clearly stops short of being truly "comprehensive" by arbitrarily downweighting and excluding items. Given that the results for each component of the composite were known beforehand and decisions seem arbitrary, we also doubt the "a priori basis" of the decisions.

Table 1. Comparing the partner attractiveness composites used by Gangestad and Dinh (2021) and our alternative composites

Item	**G&D**	**Alt**	**Scale correlation**	**Moderator t value**
Short-term (ST) attractiveness items				
ST1[b]: How difficult would it be for your partner to find another partner for a long-term relationship, who is as desirable as you?	~1/5	0	0.21	0.87
ST2: How difficult would it be for your partner to find another partner for a short affair or one-night stand, who is as desirable as you?	~1/5	0	0.42	-0.48
ST3: Other women notice my partner.	~1/5	1	0.70	-1.56
ST4: Other women feel attracted to my partner.	~1/5	1	0.71	-1.42
ST5: My partner is rarely complimented by other women.	~1/5	1	0.53	-0.27
ST6: Who do you think is more successful with members of the opposite sex [you/your partner]?	~1	0	0.41	-1.23
ST7: How satisfying is the sexual intercourse with your partner?	~1	0	0.24	-2.50/ 2.44[a]
ST8: How attractive is your partner for other women, compared to other men?	0	1	0.70	-0.18
ST9: How would you rate your partner’s desirability as a short-term mate (e.g. a partner in a one-night stand sexual encounter or brief affair) compared to other potential partners?	~1	1	0.55	-1.79
ST10: How sexy is your partner?	~1/2	1	0.68	-0.21

ST11: How physically attractive is your partner?	~1/2	1	0.60	-1.97
Long-term (LT) attractiveness items				
LT1: How would you rate your partner's desirability as a long-term mate (e.g. marriage) compared to other potential partners?	1	1	0.18	0.33
LT2: What is your partner's current financial status, compared to other potential partners?	0	1	0.76	-1.21
LT3: How would you judge your partner's likely future professional success, compared to other potential partners?	0	1	0.50	-0.25
LT4: How much money does your partner make monthly (euros)?	0	1	0.56	-0.37

Note.

The text shows the translated item, the numbers show the weight with which the items entered the composite. Weights are shown as approximate (~) when components were standardised. The scale correlation shows the corrected correlation (Revelle, 2018) with the respective composite (ST/LT). We ran individual models with each item as a moderator of the fertile window effect on the outcome extra-pair desire and behaviour and extracted the t values for the moderator test.

G&D=How this item entered the composite by Gangestad and Dinh (2021)
Alt=How this item entered our alternative composite

a First value shows the t value after mean imputation, second value after listwise deletion.

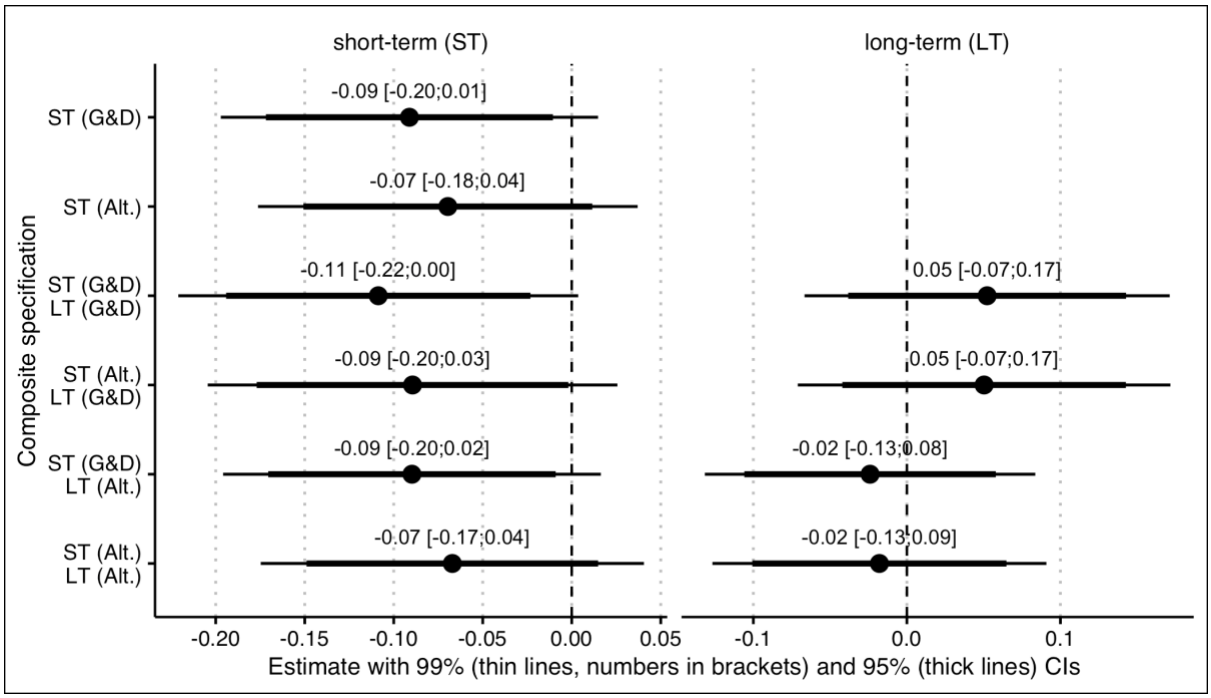
b Gangestad and Dinh (2021) lumped this item into their short-term composite, but it is clearly about long-term attractiveness. We report the scale correlation with other short-term items, but argue that the item should best be cut because of its difficult phrasing.

Paths not taken

To demonstrate the vagaries of data-dependent analysis, we show what happens when alternative decisions are made at the junctions **a)-f)** above. For the sake of simplicity, we form one alternative composite for both short- and long-term attractiveness and examine the same outcome Gangestad and Dinh (2021) focused on, extra-pair desire and behavior (EPDB). We do not standardize components, we weight items by their variance, and we exclude items that are confounded with participant's own mate value (ST1, ST2, ST6), are poorly worded (ST1, ST2), or have missing data (ST7). For short-term attractiveness, our new seven-item composite (ST3-5, ST8-11) has an internal consistency of Cronbach's α = 0.82 (95% CI [0.81;0.84]) and correlates .89 with Gangestad and Dinh's (2021) score. For long-term attractiveness, our new four-item composite (LT1-4, Table 1) has an α of 0.59 (95% CI [0.55;0.62]) and correlates .42 with the single item used both by Gangestad and Dinh (2021) and in our preregistered analyses (Arslan et al., 2018). We estimated models with our alternative composites as moderators of the fertile window effect on EPDB. In estimating these models, we allowed slopes to vary, used

the most inclusive sample size definition, and included interaction controls (see Note S2). In these specifications, the 99% confidence intervals (and most 95% confidence intervals) of the relevant moderator effect include zero (Figure 1). Of course, our alternative composites were formed knowing the data as well. But the exercise shows that arbitrary decisions, made knowing the results for each individual component, were necessary to obtain a nominally significant "joint probability of all observed effect sizes under the null hypothesis" (Gangestad & Dinh, 2021). Even though two moderators we had tested were affected by coding errors, we had also correctly reported five alternative specifications in the robustness analyses of Arslan et al. (2018), where even 95% confidence intervals included zero. Therefore, we were surprised that Gangestad and Dinh (2021) claimed "robust evidence for moderation effects," and that our data offered "the most compelling evidence for moderation effects to date" (although we agree that previously published evidence was not very compelling). To us, our evidence can be best summarised as mixed and uncertain. That the multiple moderator specifications reported here tended into the same direction for the EPDB outcome should not be overinterpreted, since the specifications stem from correlated variables within the same sample. A better measure of partner attractiveness than the one we based on items from the literature should be developed in *independent* data and, ideally, validated against a more direct measure of genetic quality, such as mutational burden scores. A measure should never be chosen because it yields desired results.

Figure 1. Estimated moderator effects on Extra-Pair Desire and Behavior using different composite specifications of Partner Attractiveness.



Note. The bars show the interaction effect of the respective composites with the fertile window effect on extra-pair desire and behaviour. ST=short-term partner

attractiveness, LT=long-term partner attractiveness. G&D=Gangestad and Dinh's (2021) measures, Alt=Our alternative composite measures.

Out-of-sample generalization

According to Gangestad and Dinh (2021), our approach in Arslan et al. (2018) did not make the best possible use of the data. We felt that our robustness analyses (Arslan et al., 2018) were fairly exhaustive, but being non-confirmatory, they should be interpreted with caution. Several techniques allow for cautious data-dependent analysis, such as cross- validation, differential privacy, and, maybe simplest, validation in a holdout (Arslan, 2017). Unfortunately, we kept no holdout for this dataset. Gangestad and Dinh's (2021) reanalysis reports more than a hundred *p*-values unrelated to our preregistered tests and makes no use of cross-validation or related techniques. Hence, we fear that Gangestad and Dinh's (2021) "robust evidence" may be random variation (Gelman & Loken, 2014) in the present sample (compare also Stern et al., 2019). In the end, we may all have to get used to the idea that we exhaust the generalizable knowledge obtainable from a dataset more easily than we think.

Fortunately, we have since collected another, very similar dataset (Arslan et al., 2020), partly in an effort to address the shortcomings of the current study with new data. Given our concerns about data-dependent analysis, out-of-sample generalizability of *post-hoc* analyses conducted by us and Gangestad and Dinh (2021) could be poor. Of course, differences between the studies in terms of heterogeneity across participants, measures, settings, and time could also affect generalizability, but by claiming "robust evidence" Gangestad and Dinh (2021) imply that slight differences should not matter. We look forward to presenting these data.

Conclusion

We are glad that errors in our analyses were corrected, that we could be frank about our underconstrained preregistration, and that we could summarise the corrected evidence in our own words. Given that the evidence for moderation in the preregistered analyses was weak at best and the results from *post-hoc* reanalyses by us and Gangestad and Dinh (2021) changed with specification decisions, we cannot concur that our data (Arslan et al., 2018) yielded "robust evidence" for GGOSH-related moderators.

However, we agree that the evidence was more mixed and uncertain than our initial brief summary (Arslan et al., 2018) made it seem. We will return to this question with new data, but will also address the lack of theoretical clarity around the GGOSH (e.g., see Note S1, S3). Dealing with the challenge of balancing overfitting, underfitting and transparency will be a crucial task for psychology as we grapple with the replication crisis. In part because of the lessons learned in this correction process, we have tried to make our procedures more resilient through several measures, such as code reviews, bug bounties (Arslan, 2018), automated testing

(Wickham, 2011), scale aggregation based on metadata (Arslan, 2019), and automated generation of components (Rouder, Haaf, & Snyder, 2019). We also increasingly conduct Registered Reports with pre-written analyses on simulated data instead of mere verbal preregistrations, which should reduce uncertainty about the planned analyses and allow outside input on best practices before data collection. We encourage researchers to share data and code (including data processing code and including data and code of older studies that are now being replicated), and to make time to vet others' code (Lakens, 2020), including ours. We hope that these procedures make future errors less likely and increase the chances that errors are detected - as has happened here. "It’s only thanks to error detectors that we can proclaim that science is self-correcting" (Vazire, 2020).

References

Arslan, R. C. (2017, September 14). Overfitting vs. Open Data. *The 100% CI*. http://www.the100.ci/2017/09/14/overfitting-vs-open-data/

Arslan, R. C. (2018, October 26). Bug Bounty Program. *One Lives Only to Make Blunders*. https://rubenarslan.github.io/bug_bounty.html

Arslan, R. C., Driebe, J. C., Stern, J., Gerlach, T. M., Ostner, J., & Penke, L. (2020). *Goettingen Ovulatory Cycle Diaries 2*. Open Science Framework. https://doi.org/10.17605/OSF.IO/D3AVF

Arslan, R. C., Schilling, K. M., Gerlach, T. M., & Penke, L. (2018). Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000208

Arslan, R. C., Schilling, K. M., Gerlach, T. M., & Penke, L. (2019). "Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior": Correction to Arslan et al. (2018). *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000251

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Gangestad, S. W., & Dinh, T. (2020). Robust Evidence for Moderation of Ovulatory Shifts by Partner Attractiveness in Arslan et al.'s (2020) Data. *Journal of Personality and Social Psychology*. https://www.researchgate.net/publication/341072189_Robust_Evidence_for_Moderation_of_Ovulatory_Shifts_by_Partner_Attractiveness_in_Arslan_et_al's_2020_Data

Gangestad, S. W., Haselton, M. G., Welling, L. L. M., Gildersleeve, K., Pillsworth, E. G., Burriss, R. P., Larson, C. M., & Puts, D. A. (2016). How valid are assessments of conception probability in ovulatory cycle research? Evaluations, recommendations, and theoretical implications. *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society*, *37*(2), 85–96. https://doi.org/10.1016/j.evolhumbehav.2015.09.001

Gangestad, S. W., Thornhill, R., & Garver, C. E. (2002). Changes in women's sexual interests and their partners' mate-retention tactics across the menstrual cycle: Evidence for shifting conflicts of interest. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1494), 975–982. https://doi.org/10.1098/rspb.2001.1952

Gelman, A., & Loken, E. (2014). The statistical crisis in science: data-dependent analysis — a 'garden of forking paths' — explains why many statistically significant comparisons don't hold up. American Scientist, 102(6), 460. https://doi.org/10.1511/2014.111.460

Harris, C. R., Chabot, A., & Mickes, L. (2013). Shifts in methodology and theory in menstrual cycle research on attraction. *Sex Roles*, *69*(9-10), 525–535. https://doi.org/10.1007/s11199-013-0302-3

Haselton, M. G., & Gangestad, S. W. (2006). Conditional expression of women's desires and men's mate guarding across the ovulatory cycle. *Hormones and Behavior*, *49*(4), 509–518. https://doi.org/10.1016/j.yhbeh.2005.10.006

Lakens, D. (2020). Pandemic researchers - recruit your own best critics. *Nature*, *581*(7807), 121. https://doi.org/10.1038/d41586-020-01392-8

Pillsworth, E. G., & Haselton, M. G. (2006). Male sexual attractiveness predicts differential ovulatory shifts in female extra-pair attraction and male mate retention. *Evolution and*

Human Behavior, *27*(4), 247–258. https://doi.org/10.1016/j.evolhumbehav.2005.10.002

Revelle, W. (2018). *psych: procedures for psychological, psychometric, and personality research* (Version 1.7.3) [Computer software]. https://CRAN.R-project.org/package=psych

Rouder, J. N., Haaf, J. M., & Snyder, H. K. (2019). Minimizing Mistakes in Psychological Science. *Advances in Methods and Practices in Psychological Science*, *2*(1), 3–11. https://doi.org/10.1177/2515245918801915

Stern, J., Arslan, R. C., Gerlach, T. M., & Penke, L. (2019). No robust evidence for cycle shifts in preferences for men's bodies in a multiverse analysis: A response to Gangestad et al. *Evolution and Human Behavior*, *40*(6), 517–525. https://doi.org/10.1016/j.evolhumbehav.2019.08.005

Vazire, S. (2020). A toast to the error detectors. *Nature*, *577*(7788), 9. https://doi.org/10.1038/d41586-019-03909-2

Wickham, H. (2011). testthat: Get started with testing. The R Journal, 3(1), 5–10.

Supplement: The evidence for good genes ovulatory shifts in Arslan et al. (2018) is mixed and uncertain

Note S1:

Gangestad and Dinh (2021) report a reanalysis of a subscale of our mate retention scale. However, their summary of (a) our previous reporting, and (b) the existing literature at the time of our preregistration, are, in our view, misleading. **a)** We never "claimed that [we] could not test moderation effects on this outcome [mate retention]". We tested moderation effects for the outcome we had preregistered, which lumped partner attentiveness and proprietariness. It is also not true that "Arslan et al. did not report the results of their exploratory analyses." (Gangestad & Dinh, 2021). They were reported in our online supplement: https://rubenarslan.github.io/ovulatory_shifts/4_stan_brms_by_item.html#male_jealousy_1 **b)** Contrary to Gangestad and Dinh (2021), the previous literature had not always reported "[minimal covariation]" between the mate retention components, rather, Gangestad et al. (2002) report "the two components [attentiveness and proprietariness] correlated substantially with one another: 0.47". Several other papers simply did not report correlations and none reported on the associations in within-subject changes across time, the relevant coefficient for our question (for comparison, between-subjects, the attentiveness and proprietariness subscales were correlated 0.25 in our data). That was the literature we based our measures and tests on. Because we found—only post-hoc—that changes in attentiveness and proprietariness did not cohere across days in the diary, we ran exploratory analyses on main effects on an item-by-item basis and summarised them as follows "Based on these analyses and research published after our preregistration (Gangestad, Garver-Apgar, Cousins, & Thornhill, 2014), future research on partner mate retention should more clearly and comprehensively examine prohibitive behaviors, as opposed to persuasive behaviors, because items measuring the former seemed to show stronger changes." (Arslan et al., 2018, p. 16). In our view, running all six moderation models in an exploratory manner for each item would be an inappropriate approach because the combinatorial explosion would make generalizable insights unlikely. We instead included an improved measure of proprietariness in our second, currently unpublished preregistered study (Arslan et al., 2020) to follow up on these unclear results in a preregistered analysis, so as not to overinterpret potential chance findings.

In their supplement section S11, Gangestad and Dinh (2021) report only the interaction effects without main effects or conditional effects. Although the interaction effects they chose to focus on are in the predicted direction, the form of the interaction is that of a crossover interaction (Widaman et al., 2012), which includes that very attractive men are reported to *decrease* in mate retention when their partners are fertile and there is no significant main effect of fertile window probability on proprietariness. We do not think this is the pattern predicted by the GGOSH; we

would expect a pattern of *attenuated increases* in mate retention, as with extra-pair desire. Given that this was a post-hoc test, we caution against overinterpreting this result.

Note S2:

In the corrected robustness analyses, reported here (Table S2), we included more data by using a continuous fertile window estimate (including more days per participant) and by relaxing exclusion criteria (after seeing that excluded women did not exhibit smaller ovulatory changes, as we had expected). We also allowed the slope of the fertile window probability to vary by participant (Barr et al., 2013) and added interaction controls for (pre-)menstruation, as advocated by Gangestad et al. (2019).

When not constrained by the preregistration, we do not think it makes sense to report models with suboptimal specifications (e.g., windowed fertile window predictors without allowing slopes to vary). Gangestad and Dinh (2021) seemed to agree on this in principle, but still presented several such models and interpreted *p*-values based on them. In Arslan et al. (2018), we had interpreted *p*-values for robustness moderator models without random slopes, but now consider doing so inappropriate. Thus, our robustness analyses, reported in Table S2, mirror Gangestad and Dinh's (2021) Table S4A, with two changes. We include in-pair desire and partner mate retention as outcomes and we include interaction controls for (pre-)menstruation. Whereas the windowed predictors exclude days close to menses, the continuous fertile window predictor is confounded with (pre-)menstruation, so these cycle phases should be adjusted for. As Gangestad et al. (2019) explain, any confound of a main effect should also be included as an interaction control when interactions are of interest. Neither Arslan et al. (2018) nor Gangestad and Dinh (2021) did so. Interaction controls make little difference to the effect sizes in this case, but explicitly include uncertainty resulting from confounding in the model.

Because the robustness analyses were not preregistered and many were run, *p*-values and confidence intervals based on these models do not have a straightforward interpretation, and it seems appropriately cautious to mentally adjust any estimates to be even more uncertain than the nominal confidence intervals would warrant.

Although the usable sample size in our robustness analyses (Table S2) was greatly increased compared to the preregistered tests (Table S1), we urge caution before a confident interpretation of the moderator analyses. Gangestad and Dinh (2021) write "the majority of women in the robustness sample were excluded from the smaller sample only because they completed fewer than 30 daily diaries, which was a preregistered exclusion criterion." This is inaccurate: we excluded these women not *only* because they did not participate for 30 days, but because they consequently never filled out the follow-up survey. Hence, among other things, we did not know whether they took hormonal medication during the study, a crucial confounder. In our robustness sample, we included women who were more likely to be anovulatory (e.g., peri-menopausal), women who had cycles longer than 37 or shorter than 22 days, and women who used hormonal medication. Estimated ovulatory changes in

these women could be attenuated. As a result, estimated main effects could be attenuated, although our robustness analyses (Arslan et al., 2018, SOM) found no strong evidence that this happened. However, if confounds, such as age, are correlated both with anovulation and with a moderator, such as partner attractiveness, it becomes more difficult to ascertain the causal role of the moderator, as we noted previously (Arslan et al., 2018, p. 4). More direct tests of ovulation seem to be a better solution to this problem than the inclusion of many additional interaction controls.

Note S3:

The theoretical predictions we tested in Arslan et al. (2018), which we labelled the GGOSH, have only been made verbally in the literature (Haselton & Gangestad, 2006). The verbal theory and the reasoning in Haselton and Gangestad (2006) are not precise enough to specify a formal model, and our preregistration shared the same flaw. Specific empirical studies have formulated specific statistical models, but these were not clearly reported and justified.

We understood GGOSH to predict at its core that women with partners who do not have good genes (GG-) should show ovulatory increases in extra-pair desire, whereas women with partners who have good genes (GG+) should not. This interpretation of GGOSH formed the basis for the majority of our preregistered moderator tests. In an elaboration of this, we also understood GGOSH to make the additional prediction that the aforementioned ovulatory increases should be restricted to women who have a providing partner (P+).

Conceptually, we think subtracting long-term from short-term attractiveness as a moderator (or adjusting for long-term attractiveness as a moderator) maps poorly onto the verbal predictions made by GGOSH. According to Gangestad and Dinh (2021), "Haselton and Gangestad (2006) and Pillsworth and Haselton (2006) previously argued for the importance of controlling for women’s ratings of their partner’s LT attractiveness (to account for possible positivity biases and scale-usage effects)", but neither study makes reference to the concepts of positivity bias or scale-usage effects. Pillsworth and Haselton (2006) reported no significant moderator effect of investment attractiveness (in their reporting, both whether or not they fit multiple moderators jointly and the direction of the effect are unclear). Haselton and Gangestad (2006) wrote "a difference score should better tap the extent to which a mate specifically has the qualities particular to good long-term mates (e.g., willingness to invest) or particular to good short-term mates (sexual attractiveness)", but in a *difference score* partners who have both "particular qualities" at the same time are penalised. Including partner long-term attractiveness as an additional moderator allows more flexibility, but we do not see how the prediction that long-term attractiveness would have an opposite effect of short-term attractiveness follows from GGOSH.

In Arslan et al. (2018), when we formulated specific statistical models, we did so with the understanding that GGOSH would predict that women who have providing partners (P+) without good genes (GG-) would show stronger ovulatory increases in extra-pair desire, whereas women who either do not have a good provider (P-), or who have a partner who both provides and supplies good genes (P+GG+) should show weaker increases. However, subtracting LT from ST tests a model where women with P+GG+ partners should show larger shifts than women with P-GG+ partners. Hence, we tested the model we thought followed from the theory (a three-way interaction between fertile window probability, ST and LT attractiveness). Gangestad and Dinh (2021) disagreed with us on this point. As alternative approaches, we included subtracting and adjusting for long-term attractiveness as two further tests in our correction (Arslan et al., 2019) and in this rejoinder. For future research on GGOSH, we recommend the simpler specification of a single moderator (short-term attractiveness), though Gangestad and Dinh (2021) seem to favour a dual moderator model (short-term and long-term attractiveness, with opposite effects). Even more preferable would be more direct measures of *good genes*, such as mutational burden scores, instead of purported proxies like short-term attractiveness that may additionally suffer from "positivity bias" and "scale-usage effects" (Gangestad and Dinh, 2021).

Table S1: The preregistered moderation tests after corrections (141 women across 1915 days).

Outcome	Specification	Term	Estimate [99% CI]	*p*-value
Extra-pair desire and behaviour	Physical Attractiveness		-0.06 [-0.22;0.11]	0.395
	ST Attractiveness		-0.08 [-0.26;0.09]	0.212
	ST x LT Attractiveness	ST	-0.09 [-0.27;0.10]	0.216
		LT	0.03 [-0.14;0.21]	0.636
		ST x LT	0.01 [-0.14;0.16]	0.860
	ST - LT Attractiveness		-0.07 [-0.23;0.09]	0.253
	ST Attractiveness w/ LT controlled	ST	-0.09 [-0.27;0.09]	0.180
		LT	0.03 [-0.14;0.21]	0.641
	Partner Attractiveness vs. Own		-0.07 [-0.24;0.10]	0.274
In-pair desire	Physical Attractiveness		-0.21 [-0.53;0.12]	0.104
	ST Attractiveness		-0.24 [-0.58;0.09]	0.062
	ST x LT Attractiveness	ST	-0.35 [-0.71;0.00]	0.011
		LT	0.14 [-0.20;0.48]	0.284
		ST x LT	-0.25 [-0.53;0.04]	0.025
	ST - LT Attractiveness		-0.24 [-0.56;0.07]	0.046
	ST Attractiveness w/ LT controlled	ST	-0.28 [-0.63;0.07]	0.037
		LT	0.16 [-0.18;0.50]	0.236
	Partner Attractiveness vs. Own		-0.09 [-0.42;0.24]	0.503
Partner mate retention	Physical Attractiveness		-0.03 [-0.26;0.21]	0.776
	ST Attractiveness		-0.02 [-0.25;0.22]	0.869
	ST x LT Attractiveness	ST	-0.06 [-0.31;0.20]	0.564
		LT	0.05 [-0.20;0.29]	0.622
		ST x LT	-0.10 [-0.31;0.10]	0.192
	ST - LT Attractiveness		-0.05 [-0.27;0.18]	0.605
	ST Attractiveness w/ LT controlled	ST	-0.03 [-0.27;0.22]	0.775
		LT	0.05 [-0.19;0.30]	0.574
	Partner Attractiveness vs. Own		-0.11 [-0.35;0.12]	0.212

Note. In these analyses, the aggregation of the *Partner Attractiveness vs. Own* and the *ST attractiveness* variable moderators were corrected (by correcting the jumbled order of items for relative attractiveness and by imputing the mean for missing values in sexual satisfaction, respectively). The column *Specification* refers to how each moderation model was specified. In two specifications, both short- (ST) and long-term (LT) attractiveness were entered as moderators of the

fertile window effect, so the *Term* column disambiguates the coefficients for each. For the other models, the specification refers to a single moderator.

As in Arslan et al. (2018) but not as in Gangestad and Dinh (2021), fertile window probability estimates are not standardised, so moderator effects are interpretable as changes to the effect of fertile window probability. Some *p*-values do not match down to the second digit with Gangestad and Dinh (2021), because they standardized moderator variables at level 2 (woman) as if they were on level 1 (diary days), that is, the standard deviation they computed was slightly incorrect because women contributed different numbers of days to the diary.

Table S2: The corrected and improved robustness analyses of moderation (429 women across 10,395 days).

Outcome	Specification	Term	Estimate	[99% CI]	[95% CI]
Extra-pair desire and behaviour	Physical Attractiveness		-0.06	[-0.17;0.06]	[-0.14;0.03]
	ST Attractiveness		-0.11	[-0.22;0.01]	[-0.19;-0.02]
	ST x LT Attractiveness	ST	-0.13	[-0.25;-0.00]	[-0.22;-0.03]
		LT	0.06	[-0.08;0.19]	[-0.05;0.16]
		ST x LT	-0.01	[-0.11;0.09]	[-0.09;0.07]
	ST - LT Attractiveness		-0.11	[-0.23;0.01]	[-0.20;-0.02]
	ST Attractiveness w/ LT controlled	ST	-0.13	[-0.25;-0.00]	[-0.22;-0.03]
		LT	0.06	[-0.07;0.19]	[-0.04;0.16]
	Partner Attractiveness vs. Own		-0.06	[-0.17;0.06]	[-0.14;0.03]
In-pair desire	Physical Attractiveness		0.05	[-0.17;0.27]	[-0.12;0.22]
	ST Attractiveness		-0.00	[-0.22;0.22]	[-0.17;0.16]
	ST x LT Attractiveness	ST	0.02	[-0.22;0.26]	[-0.17;0.20]
		LT	-0.02	[-0.28;0.24]	[-0.22;0.17]
		ST x LT	-0.03	[-0.22;0.17]	[-0.18;0.12]
	ST - LT Attractiveness		0.01	[-0.23;0.25]	[-0.17;0.19]
	ST Attractiveness w/ LT controlled	ST	0.01	[-0.23;0.25]	[-0.18;0.19]
		LT	-0.02	[-0.26;0.23]	[-0.20;0.17]
	Partner Attractiveness vs. Own		-0.06	[-0.28;0.16]	[-0.22;0.11]
Partner mate retention	Physical Attractiveness		-0.01	[-0.16;0.14]	[-0.12;0.10]
	ST Attractiveness		0.01	[-0.14;0.16]	[-0.10;0.12]
	ST x LT Attractiveness	ST	0.03	[-0.14;0.19]	[-0.10;0.15]
		LT	-0.05	[-0.23;0.12]	[-0.19;0.08]
		ST x LT	-0.03	[-0.17;0.10]	[-0.13;0.07]
	ST - LT Attractiveness		0.04	[-0.12;0.20]	[-0.08;0.16]
	ST Attractiveness w/ LT controlled	ST	0.03	[-0.14;0.19]	[-0.10;0.15]
		LT	-0.04	[-0.21;0.13]	[-0.17;0.09]
	Partner Attractiveness vs. Own		-0.12	[-0.27;0.02]	[-0.23;-0.01]

Note. This table can be read the same as Table S1. These models were run on the largest usable sample of women not on hormonal contraception. Because these models implement several best practices (see Note S2) that deviate from our preregistration, they are presented without p values.

Figure S1:

Moderation for an ovulatory shift model on male proprietariness, without adjusting for long-term attractiveness. The moderator is Gangestad and Dinh's, (2021) partner attractiveness composite. Dots show the raw data in each moderator quintile (jittered and transparent to reduce overplotting). Lines show the model-estimated marginal effect of the fertile window variable mid-quintile with 95% CIs. Color reflects the moderator values. Rather than showing the expected attenuated effect for above-average partners, the slope turns negative in the upper quintiles, that is, attractive men are *less* proprietary when their (naturally cycling) partners are in the fertile window.

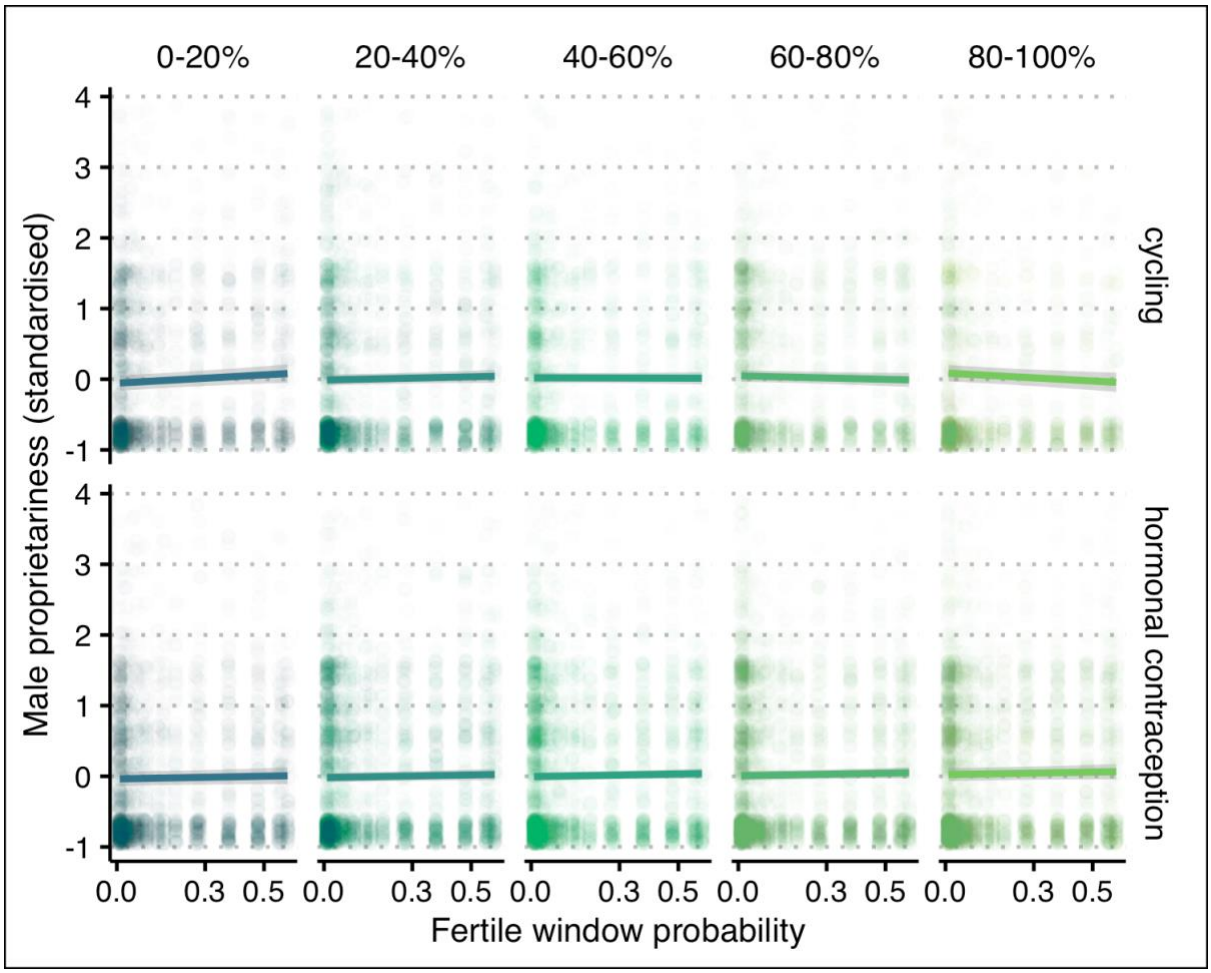


Table S3. Comparing and contrasting Gangestad and Dinh's (2021) account with our own account. Although we agree with many of the criticisms raised by Gangestad and Dinh (2021), in some instances, they do not accurately summarise our own reporting and conclusions. In this table, we compare their summaries with quotes from our paper and our correction and give our own summary.

Gangestad & Dinh (2021)	Arslan et al. (2018/2019)	Our summary
"In their published report, Arslan et al. did not acknowledge their preregistered α of .05." Regarding our power analysis: "This target sample size implies α = .05." (p. X)	Arslan et al. 2018 (p. 12): "Because we had not preregistered a procedure to correct for multiple comparisons due to multiple outcomes and believed Bonferroni to be too conservative, as many outcomes were highly correlated, we tested whether we would have ever rejected the null hypothesis of no effect in our HC control group with the significance threshold of .01. Although this would have been the case for one outcome, follow-up analyses showed that this result would not have survived our robustness analyses, so we concluded that our chosen threshold was appropriate. The pattern of significant results here would not have been different using the uncorrected threshold of .05 or when using a Benjamini-Hochberg correction (Benjamini & Hochberg, 1995; see	There was no need to infer an α from our power analysis. We clearly acknowledged that we had preregistered a conventional alpha threshold but no procedure to correct for multiple comparisons. We were explicit about our reasoning to adopt a more stringent threshold, which we still think is sound. Gangestad and Dinh (2021) make no convincing case why an uncorrected threshold would be appropriate.

	supportive website, osf.io/pbef2)."	
"Arslan et al. tested their hypotheses in samples using several sets of criteria, none of which precisely conformed to their preregistered criteria." (p. X).	Arslan et al. 2018 (p. 7): "We preregistered several exclusion criteria that we deemed useful to exclude women with potentially anovulatory cycles, but also wrote that we would examine the effect of applying these criteria. Applying the strictest criteria proved to be overexclusive, as only 13% of the naturally cycling sample would have been retained. Hence, we differentiated our exclusion criteria into four strictness levels and examined the effect of applying these levels in robustness checks. The participant flow and exclusion criteria are shown in Figure 1."	We should have been clearer that the preregistration had two sets of criteria (from the first version and from the amendment on May 10, 2014 prior to data analysis) and that our differentiation was not exactly along those lines. However, we were transparent that our preregistered criteria were overexclusive and that we differentiated them post-hoc. We clearly labelled the criterion used for preregistered analyses as "lax". We especially regretted the criterion on cycle regularity as women were not confident in their reported regularity, so relying on this criterion might have excluded many women with regular cycles. We also decided to retain women who had broken up with their partner in the main preregistered analysis, because we thought excluding them might mean excluding the women with the strongest extra-pair desires. The decisions to differentiate the criteria like this were made before all data were

		collected and not conditional on results for ovulatory shifts.
"In their commentary in response to corrections, Arslan et al. argue that the presence of some “non-significant” effects, even with evidence for other “significant” effects, justifies their conclusion that they could not replicate previously reported moderators. The reasoning behind this argument relies on strict dichotomous judgments—significant vs. non-significant—as criteria of whether data yield evidence for or against an effect." (p. X)	Online extended correction, 2018: "Models with varying slopes indeed fit better for all outcomes. We reported robustness checks with varying slopes for all main effects, but we had not done so for our moderators tests, because we found no evidence of moderation and the check would have only made the test more conservative. Given that correcting the error led to a nominally significant result, we also tested a model allowing for slopes to vary. In this model, the predicted interaction was non-significant for extra-pair desire ($p = 0.085$). The predicted interaction for partner mate retention in the robustness check would have been significant ($p = 0.0072$) according to our threshold of .01 for the preregistered tests, but still potentially consistent with sampling error given that 24 moderator effects had been tested (four moderators, three outcomes, two subsamples) were tested	Our reasoning relied on recognising the potential for overfitting and false positives/overestimation of effects when multiple tests are carried out. It was not a "strict dichotomous judgment" but a result "potentially consistent with sampling error". We never used the phrase "evidence against an effect".

	for essentially one hypothesis."	
"In other words, Arslan et al. saw no need to alter or qualify the previous statements they made in their article regarding the purported lack of evidence they found for moderation effects." (p. X)	Online extended correction, 2018: "Overall, as we had already stressed in our discussion, it would be premature to conclude an absence of moderation: confidence intervals were too wide to rule out potentially relevant effect sizes and patterns were often in the predicted form for extra-pair desire (but not for in-pair desire). But neither should these models, which were suggested after seeing the results for other models, be seen as evidence for moderation, given the number of tests performed. If a prediction from the literature is supported in preregistered tests, checks like ours can show robustness to relaxing or tightening assumptions. The evidence for the predicted moderators is clearly not robust in our data. More data is needed to reach adequate power for more informative tests of moderation patterns, and is indeed forthcoming. Maybe more importantly, theories need to be clearer, so that they can	We now agree that our original conclusions did not hedge sufficiently. On rereading our own conclusion in the published paper, we understand why Gangestad and Dinh (2021) did not find these sufficiently hedged. Still, in our extended correction, we stressed the large uncertainty about moderation effects, not their absence, and (mistakenly) said we had been clear about this in the paper.

specify severe tests. We found this difficult to do at the time of planning the study."

Supplementary References

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

Gangestad, S. W., Garver-Apgar, C. E., Cousins, A. J., & Thornhill, R. (2014). Intersexual conflict across women's ovulatory cycle. *Evolution and Human Behavior*, *35*(4), 302–308. https://doi.org/10.1016/j.evolhumbehav.2014.02.012

Gangestad, S. W., Thornhill, R., & Garver, C. E. (2002). Changes in women's sexual interests and their partners' mate-retention tactics across the menstrual cycle: Evidence for shifting conflicts of interest. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1494), 975–982. https://doi.org/10.1098/rspb.2001.1952

Widaman, K. F., Helm, J. L., Castro-Schilo, L., Pluess, M., Stallings, M. C., & Belsky, J. (2012). Distinguishing ordinal and disordinal interactions. *Psychological Methods*, *17*(4), 615–622. https://doi.org/10.1037/a0030003