Correction to Arslan et al. (2019). Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior

true

06-20-2019

We (me, Katharina Schilling, Tanja M. Gerlach, & Lars Penke) recently published a diary study on ovulatory changes in the Journal of Personality and Social Psychology.

Unfortunately, we made a few mistakes in reporting the study. The correction is due to appear today in JPSP. In our opinion, the mistakes, although annoying and preventable, changed nothing substantive. I have taken to adding automated testing to my data cleaning code and instituted a bug bounty policy to reduce the odds of such errors in my future work.

Because corrections have to be quite short, we will use this blog post to give a little more detail. We also apologise for the time it took for the correction. Partly, this resulted from misunderstandings and partly simply from the time and number of emails it took to formulate the correction.

After expanding on the correction, we will also respond to some criticisms that we do not think are errors in our work, but differences in interpretation. Still, we want to respond to all criticism made fully and hope this can refute allegations that we wanted to deceive readers.

Figure 1 and the case numbers

We regret the following errors and inconsistencies in our published paper. Between our initial submission and our revision, we had made a small adjustment to the code for our exclusion criteria and neglected to update Figure 1 and Table 3 (because we did not notice that we had a few more participants and days). This led us to report an incorrect, lower number of total participants (1043 instead of 1054) for the robustness checks. The number of days were also off by a few hundred, as well as various sample means. The substantive results (model coefficients etc.) were reported correctly and with correct case numbers (in the online supplement).

The preregistered work is unaffected by this error. A corrected Figure 1 also shows two exclusion criteria (hypothesis guessing and long interruptions of the diary) that were mentioned on the supplementary website, but missing from Figure 1. A corrected figure can be found at http://rubenarslan.github.io/ovulatory_shifts/2_descriptives.html and in the updated article.

`effsize` package bug

An error in the `effsize` R package led to the reporting of inflated effect sizes for the Hedges' g differences between hormonal contraceptive users and non-users in Table 1. After re-analysing data for the correction, we suddenly got different effect sizes. It turned out there was a bug in the `effsize` package for Hedges' g computation that had been fixed in a newer version. On the positive side, these means we reported larger differences between our naturally cycling group and our hormonal contraception quasi-control group.

Programming error for a moderator variable

We made a programming error when aggregating the variable "partner's attractiveness relative to self". Specifically, values were accidentally sorted before aggregation, leading to nonsense ranks. Fixing this error

led to the following changes:

- In the preregistered analyses, the moderation of fertile window effects on extra-pair desire and behaviour was no longer non-significant in the opposite direction of the prediction, but non-significant in the predicted direction ($p = 0.23$).
- In the robustness analyses, the predicted interaction was significant for extra-pair desire and behaviour ($p = 0.00565$) and partner mate retention ($p = 0.0014$).

Our preregistered tests, following the literature at the time, had not permitted slopes for menstruation and the fertile window to vary by woman, even though fitting a cross-level moderation essentially stipulates that varying slopes must exist.

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. This model rendered the predicted interaction 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 predicted interactions effect (four moderators, three outcomes, two subsamples) were tested for essentially one hypothesis.

Other post-publication feedback (not part of the correction)

Figure 5

Dan Engber helpfully pointed out that the caption for Figure 5 was not sufficiently clear. The figure was intended to show differences in patterns across the cycle. To this end, we standardised differences within variable and hormonal contraceptive status ("within-subject change" in the figure caption). This focuses the eye on the differences in changes for HC users and non-users. In Figure 3, we also showed the mean differences. An alternative version of Figure 5 can be found online: http://rubenarslan.github.io/ovulatory_shifts/3_stan_brms_long2.html

Following the preregistration

A reader accused us of deception for not following our preregistration to the letter. It was our intention to be faithful to the preregistration as much as possible and transparent about any deviations. We think we succeeded in doing so and that problems raised by the reader are mainly problems of explicitness and interpretation. It was our first preregistration (in 2014), we had no models to go on how to preregister correlational work with many simultaenous (but not all related) hypothesis tests. It was also our first menstrual cycle study. For this reason, we relied on expert opinion (including the aforementioned reader) to design, for example, our exclusion criteria.

This process led to a few suboptimalities (an incomplete list, I am sure):

- Our exclusion criteria were overly strict and would have led to excluding most of the women for no reason (according to effect size estimates, excluded women were not more likely to be anovulatory).
- We preregistered the use of windowed fertility predictors, which throw away most of the variation and days.
- We preregistered no strategy to deal with multiple testing, although we had multiple outcomes (some of which were highly correlated).
- We preregistered several moderators that were all designed to test the same hypothesis, instead of the strongest possible specification.
- We did not preregister how we would aggregate some of the more complex items in the data.

We think we were transparent about how we chose to deal with these problems. We definitely did not make any decisions to arrive at foregone conclusions, instead, we think we had good reasons for non-arbitrary decisions.

Operationalisation of hypothesis 2.2

The reader alerted us that our hypothesis ***H.2.2.***[1] could also be interpreted to mean a different statistical model than the one we fitted.

We interpreted it as meaning that women who have a partner who is high in long-term attractiveness but low in short-term attractiveness would show ovulatory increases in extra-pair desire, whereas all other women would not. Basically, women who have a partner who is a "provider" but does not have "good genes" would be interested in extra-pair men; other women would not be.

We saw this in contrast to the simpler model, which we also fit, with only short-term attractiveness as the moderator. The reader interpreted it as meaning that we should adjust for long-term attractiveness to remove a "positivity bias" and test only the interaction between the fertile window and short-term attractiveness. Previous work had sometimes tested such a model and sometimes a difference score.

Although we reported them, we recommend not interpreting difference scores such as this (or the relative attractiveness variable above) in isolation, because they assume that women with partners who are attractive for both long- and short-term relationships behave the same way as women with partners who are not attractive for either long- or short-term relationships. We think this is not what the verbally specified theory predicts, but of course verbal specifications can be debated because they often leave some room for ambiguity.

In our preregistered analyses, none of these alternative specifications would have yielded a significant effect, except one significant result in the opposite direction for in-pair desire. However, in our robustness checks, the interaction for this alternative specification would have been significant ($p = 0.006$). Again, allowing for slopes to vary rendered this interaction nonsignificant at .01 ($p = 0.045$).

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

Operationalisation of preregistration regarding hormonal contraceptive users

Lastly, we did not pre-register that we would use hormonal contraception (HC) users as a quasi-control group for the naturally cycling group. Consistent with this, our preregistered tests compared fertile window changes with zero, not with the baseline change for HC users. However, we reported the latter comparison as well, in the preregistered analysis section, because we considered our omission of a strategy against multiple testing problems a flaw in our preregistration. We thought reporting the quasi-control group was one way to show that our ad-hoc strategy was effective.

We hope these additional tests, which were in fact always consistent with our preregistered tests, did not lead to confusion regarding our preregistration. The choice of additionally presenting these analyses did not affect our conclusions and was not made conditional on the results.

[1] Moderation or shift hypotheses: The ovulatory increase in women's extra-pair desires and reported male mate retention behavior is strongest (and the in-pair desire increase is weakest) for women who perceive their partners as low in sexual attractiveness relative to long-term partner attractiveness.