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 We instead included an improved measure of proprietariness in our second, currently unpublished preregistered study (Arslan et al., 2020) to follow up on these unclear results in a preregistered analysis, so as not to overinterpret potential chance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.

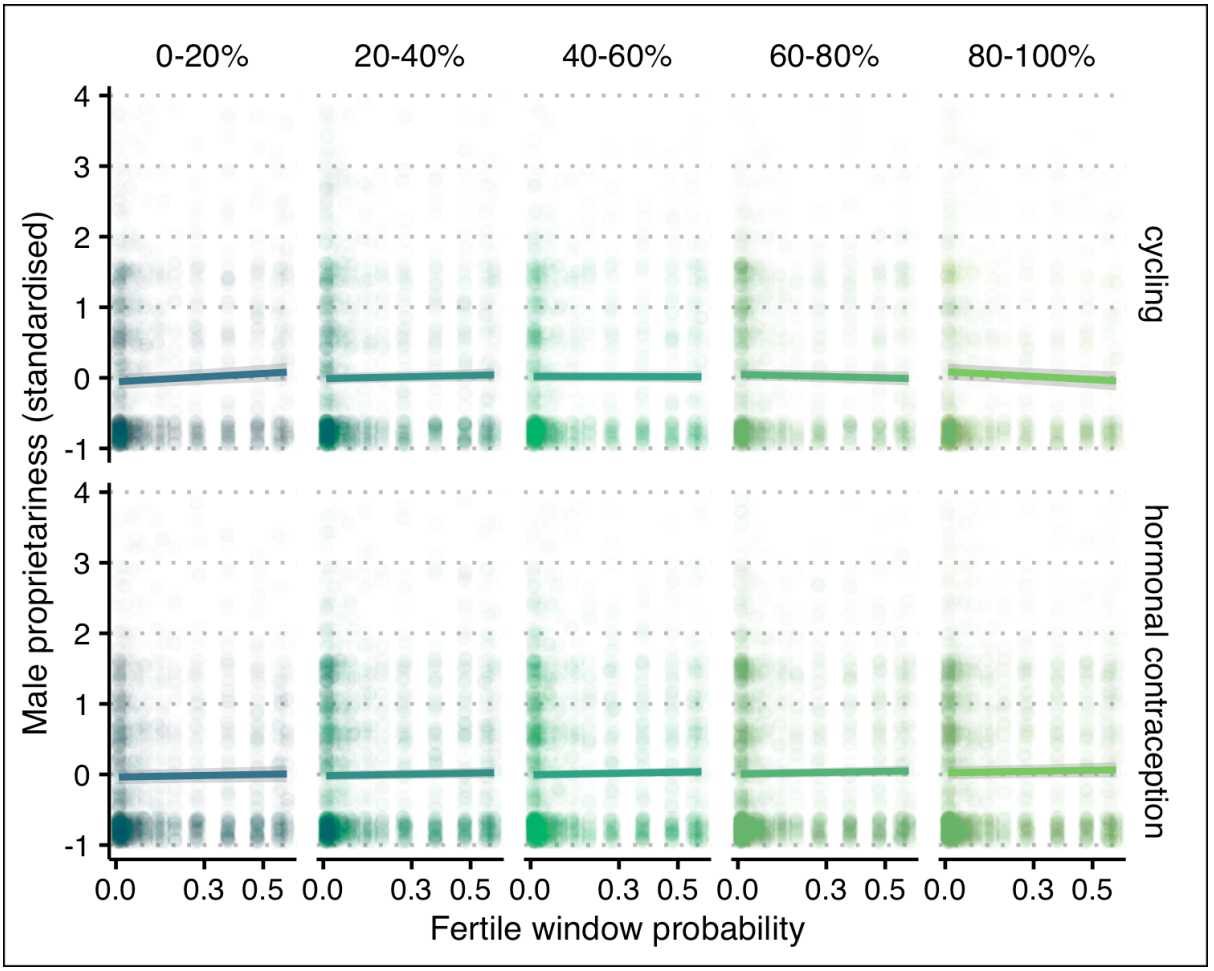

Table S3. Comparing and contrasting Gangestad and Dinh's (2021) account with our own account. Although we agree with many of the criticisms raised by Gangestad and Dinh (2021), in some instances, they do not accurately summarise our own reporting and conclusions. In this table, we compare their summaries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

	.05 or when using a Benjamini-Hochberg correction (Benjamini & Hochberg, 1995; see 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	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".

	still potentially consistent with sampling error given that 24 moderator effects had been tested (four moderators, three outcomes, two subsamples) were tested 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	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.

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."

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