Robust Evidence for Moderation of Ovulatory Shifts by Partner Attractiveness in Arslan et al.'s (2020) Data

Steven W. Gangestad & Tran Dinh

Supplementary Online Materials (SOM)

Table of Contents

For R script and output, visit https://osf.io/279ph/

S1. Descriptions of variables in Arslan et al.'s online mixed model regression tables

Arslan et al. posted corrected tables for two analyses on their online resource.

https://rubenarslan.github.io/ovulatory_shifts/3_fertility_robustness.html#partners-short-term-vs-long-term-attractiveness; https://rubenarslan.github.io/ovulatory_shifts/3_fertility_robustness.html#partners-short-term-vs-long-term-attractiveness-1

We describe the variables used in these analyses.

menstruationpre – a dummy variable controlling for the diary day being in a period of several days prior to menses

menstruationyes – a dummy variable controlling for experiencing menses on the diary day

fertile_mean – mean continuous fertility value for all diary days completed by a participant

fertile – continuous fertility value estimated for a given day based on day of the cycle and actual or anticipated first day of the next cycle

partner_attractiveness_longterm – participant's rating of partner's attractiveness as a long-term partner

partner_attractiveness_shortterm – participant's rating of partner's attractiveness as a short-term partner (composite of 3 components)

includedhorm_contra – dummy variable specifying whether women were normally cycling or using a hormonal contraceptive. Importantly, this variable was coded 0 for normally cycling women. All interactions including this variable were entered. Hence, all effects NOT including this variable, including the critical partner_attractiveness_shortterm by fertile interaction (or, in the second analysis, the critical partner_st_vs_lt by fertile interaction), are estimated for normally cycling women.

partner_st_vs_lt – a difference score of participants' ratings of partner's short-term attractiveness and long-term attractiveness; used in analysis using measure of partner short-term vs. long-term mate attractiveness as a moderator

S2. Power considerations

In their preregistration, Straus et al. (2014) estimated 80% power to detect an effect size equivalent to r = .2, with *N* = 200 and α = .05 (though not specified as a main effect or moderator effect). In fact, and as noted by a reviewer, the mixed model nature of the design makes power less-than-straightforward to assess. We can nonetheless address how much power Arslan et al. had to detect a moderator effect size observed in the robustness sample of 429. We target partner ST attractiveness as the predictor (our corrected measure). As we describe in the main text and Table S8, this effect is of meaningful size. The effect size yielded a *t*-value in the sample (-3.05) close to the expectation of *t* if power were 86% with α = .05. Arslan et al.'s α of .01, however, reduces power to 68%. With random slopes added, effect size remains near the same but power reduces to 66% and 42% in these scenarios, respectively (based on observed standard errors).

In the samples of 143 and 123 in preregistered tests (using narrow and broad windows), power is even lower. Based on relative standard errors across analyses, we estimate power of 29% and 33% power to detect these effects in the sample of 143 using narrow and broad windows, respectively, with α = .05 and no random slopes. With α = .01, power becomes 12% and 15%. With α = .01 and random slopes for fertility modeled, power becomes 8% and 9%. Arslan et al. had very little chance to detect these meaningful effects in their purported preregistered tests with α = .01, such that the absence of "significant" fertility × partner attractiveness interaction emerging from these analyses cannot speak to whether true meaningful effects exist.

One reason for modest power, despite large sample size and a highly meaningful moderation effect, is that the interaction is a spreading interaction—where the effect of fertility for women with attractive partners is smaller, but not in the opposite direction, than the fertility effect for women with relatively unattractive partners. These kinds of interaction effects require large sample sizes to detect, typically much more than researchers expect (e.g., see online post by Simonsohn at http://datacolada.org/17). As Arslan et al. also acknowledge, their methods for assessing fertility status possess moderate validity. Power in their design derives from large sample size and dense sampling, not high validity of measurement.

S3. Effect sizes, test-statistics, and *p*-values for fertility × partner attractiveness interactions: Arslan et al.'s online resource

Table 1, main text, gives the effect sizes, test-statistics and *p*-values for the fertility ×partner attractiveness interaction effects on EP interests in the robustness sample with Arslan et al.'s partner ST attractiveness measure corrected. Table S3 gives these statistics taken directly from Arslan et al.'s online resource (after correction of errors posted in June 2018 but not correction of errors in partner ST attractiveness). See https://rubenarslan.github.io/ovulatory_shifts/3_fertility_robustness.html#moderators.

Moderator	γ	**95% CI**	t	p
Partner Attractiveness, uncontrolled				
Physical Attractiveness	-0.010	-.021 to .001	-1.73	0.089
ST Attractiveness	-0.011	-.022 to .000	-1.93	0.053
Partner ST Attractiveness vs. LT				
ST – LT Attractiveness	-0.018	-.030 to -.006	-2.96	0.003
ST Attractiveness w/ LT controlled	-0.017	-.029 to -.005	-2.74	0.006
Partner Attractiveness, Relative to Self				
Partner Attractiveness vs. Own	-0.016	-.027 to -.005	-2.77	0.006

ST = short-term; LT = long-term attractiveness. *N* of normally cycling women = 429.

Effect size estimates are taken from analyses we ran using z-scored predictors. Consistent with Arslan et al.'s primary analyses, random slopes for fertility status were not modeled in these analyses.

S4. Additional material on correcting Arslan et al.'s calculation of partner ST attractiveness

The coding error

Arslan et al. provide R code they used to calculate composite variables. See https://rubenarslan.github.io/ovulatory_shifts/1_wrangle_data.html. The code they used to calculate partner ST attractiveness ('partner_attractiveness_shortterm' in their data set) follows:

```
xsection$partner_attractiveness_shortterm = scale(rowSums(xsection %>% select
(attractiveness_stp, partner_attractiveness_physical, satisfaction_sexual_int
ercourse), na.rm = TRUE)) # missings in satisfaction_sexual_intercourse for t
hose who haven't yet had sex with partner
```

where 'partner_attractiveness_physical' is the sum of two items, 'attractiveness_body' and 'sexy.' The note in the code acknowledges that there are cases missing a rating in satisfaction_sexual_intercourse. Among naturally cycling women in the robustness sample, 7% are missing a rating on this item. The code 'na.rm = TRUE' removes the '*NA*' designation for these missing values. As a result, for women who did not respond to the sexual satisfaction item, their ST attractiveness value is simply the sum of the other two components. I.e., their rating on satisfaction_sexual_intercourse is treated as though it were zero.

Our solutions to the error

Arslan et al. describe partner ST attractiveness as a composite of the three components described above: physical attractiveness, an explicit rating of the partner's attractiveness as a short-term partner (specifically, a one-night stand partner), and sexual satisfaction of the partner. The most straightforward way to create a composite these components is to unit-weight and zero-center each one through z-scoring and then to average the three z-scores:

(a) diary$STattractiveness <- rowMeans(diary[,c("Zattractiveness_stp", "Zpartner_attractiveness_physical","Zsatisfaction_sexual_intercourse")], na.rm = TRUE)

The code 'na.rm = TRUE' entails that partner ST attractiveness for women missing a rating of sexual satisfaction is the mean of the remaining two z-scored components.

In our primary analyses, we used this recoding.

A second solution is to use Arslan et al.'s sum, but remove all women who did not respond to the sexual satisfaction item:

(b) diary$partner_attractiveness_shortterm = scale(rowSums(diary[,c("attractiveness_stp", "partner_attractiveness_physical", "satisfaction_sexual_intercourse")], na.rm = FALSE))

In addition to removing all participants who did not respond to satisfaction_sexual_intercourse, (b) retains the weighting of components in Arslan et al.'s measure. Because partner_attractiveness_physical is itself a sum of responses to two items, its standard deviation is greater than that of other components, which means it receives greater weight than the other two components. By contrast, all components are weighted equally in (a). Nonetheless, the two measures strongly covary: Within normally ovulating women in the robustness sample, *r* = .99.

In this supplement, we also report analyses that use this second solution (see below).

Analyses that model random slope variation for fertility: Fertility × partner attractiveness interactions on EP interests

Arslan et al. did not model random slope variation for fertility in their primary analyses. Arguably, however, random slope variation for fertility should be modeled. Omitting it when the association between outcomes and fertility vary across women leads to alpha-inflation in tests of fixed effects involving the fertility term, including fertility × partner attractiveness interactions, largely because the error term for such fixed effects is underestimated (e.g., Barr, 2013). Random slope variation for level-1 control variables (in the case of Arslan et al.'s models, menses and premenses) are best not modeled, as their exclusion does not inflate alpha but may reduce power (Barr, 2013; Barr et al., 2013). In Table 1, main text, we report fertility × partner attractiveness interactions from analyses that do not model random slope variation for fertility, which parallel Arslan et al.'s primary analyses. In Table S4a below, we report fertility × partner attractiveness interactions from analyses that do model random slope variation for fertility. In these analyses, partner ST attractiveness is computed by averaging z-scores of its three components (formula (a) above).

Table S4a. Effect sizes, test statistics, and p-values for moderation of associations of fertility status with female extra-pair sexual interests by partner attractiveness within the robustness sample: Random slope variation for fertility modeled

Moderator	*γ*	**95% CI**	*t*	***p***
Partner Attractiveness, uncontrolled				
Physical Attractiveness	-0.011	-.026 to .004	-1.47	0.141
ST Attractiveness	-0.018	-.033 to -.003	-2.36	0.018
Partner ST Attractiveness vs. LT				
ST – LT Attractiveness	-0.025	-.041 to -.009	-3.11	0.002
ST Attractiveness w/ LT controlled	-0.026	-.042 to -.009	-3.08	0.002
Partner Attractiveness, Relative to Self				
Partner Attractiveness vs. Own	-0.016	-.027 to -.005	-2.00	0.046

ST = short-term; LT = long-term attractiveness. *N* of normally cycling women = 429.
Full results of all analyses provided in SOM, R output. Random slopes for fertility are modeled in these analyses.

Analyses using Arslan et al.'s measure of ST attractiveness but with women not responding to the sexual satisfaction item excluded: Fertility × partner attractiveness interactions on EP interests

As described above, we also ran analyses that used Arslan et al.'s measure of partner ST attractiveness, but excluded women who did not respond to all items (i.e., women who did not respond to the sexual satisfaction item). Table S4b reports analyses that do not model random slope variation for fertility. Table S4c reports analyses that do model random slope variation for fertility. As can be seen, results using this alternative solution yield results that are very similar to results reported in Table 1, main text, and Table S4a.

Table S4b. Effect sizes, test statistics, and p-values for moderation of associations of fertility status with female extra-pair sexual interests by partner attractiveness within the robustness sample

Moderator				
	γ	**95% CI**	t	p
ST Attractiveness	-0.017	-.029 to -.004	-2.75	0.006
ST – LT Attractiveness	-0.023	-.036 to -.011	-3.61	<.001
ST Attractiveness w/ LT controlled	-0.023	-.036 to -.010	-3.50	<.001

ST = short-term; LT = long-term attractiveness. *N* of normally cycling women = 429.
Full results of all analyses provided in SOM. Random slopes for fertility are not modeled in these analyses.

Table S4c. Effect sizes, test statistics, and p-values for moderation of associations of fertility status with female extra-pair sexual interests by partner attractiveness within the robustness sample: Random slope variation for fertility modeled

Moderator				
	γ	**95% CI**	t	p
ST Attractiveness	-0.018	-.034 to -.002	-2.19	0.028
ST – LT Attractiveness	-0.024	-.040 to -.007	-2.77	0.006
ST Attractiveness w/ LT controlled	-0.024	-.041 to -.007	-2.78	0.006

ST = short-term; LT = long-term attractiveness. *N* of normally cycling women = 429.
Full results of all analyses provided in SOM. Random slopes for fertility are modeled in these analyses.

S5. Analyses using narrow and broad fertility windows

Arslan et al. utilized three fertility measures: a continuous measure of conception risk (e.g., Gangestad et al., 2016), a narrow window of the fertile phase, and a broad window of the fertile phase. The latter two measures were preregistered. Yet Arslan et al. acknowledge that the continuous measure "implemented the best practices that were published after we preregistered" (p. xxx). As well, narrow and broad windows eliminate many observations relative to the continuous measure, which inflates standard errors. (See SOM, S4 on statistical power in these analyses.)

One can nonetheless examine partner attractiveness × fertility effects on extra-pair sexual interests and confidence intervals using Arslan et al's narrow and fertile windows. We ran all analyses on the 143 women within Arslan et al.'s "lax" sample, as well as the subset of 123 women who met Arslan et al.'s preregistered criteria. See Table S5 below. We also report effects in analyses on these samples using the continuous fertility measure. All 36 effects run in a negative direction. The continuous measure yields 11 $p < .05$ out of 12 tests. Narrow and broad windows yield 5 $p < .05$ out of 24 tests but, once again, one should not infer null effects from binary "non-significant" results, particularly when power is low (see SOM S2).

Table S5. Effect sizes, test statistics, and p-values for moderation of associations of fertility status with female extra-pair sexual interests by partner attractiveness within the samples of 143 and 123

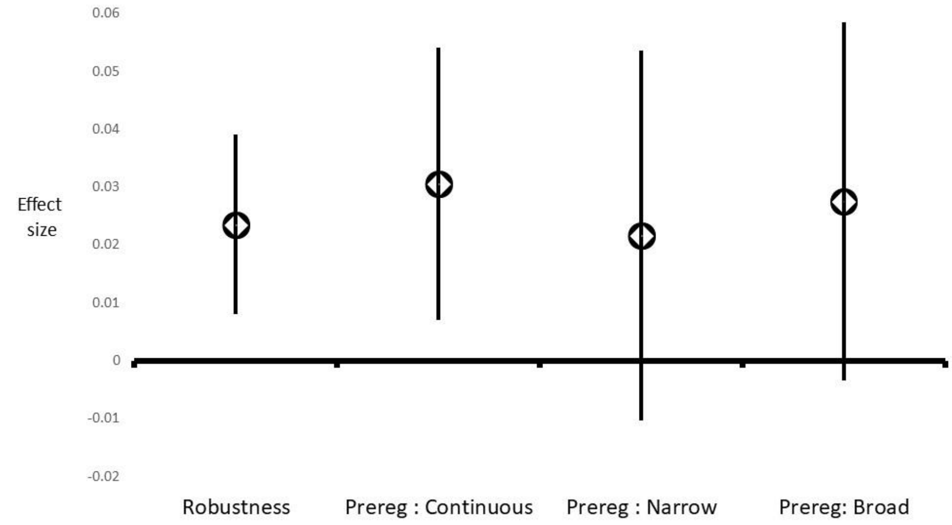
Sample of 143	Continuous				Narrow Windows				Broad Windows					Robustness[1]
	γ	SE	*t*	*p*	γ	SE	*t*	*p*	γ	SE	*t*	*p*		γ
Physical Attractiveness	-0.022	0.009	-2.52	0.012	-0.011	0.012	-0.93	0.354	-0.004	0.011	-0.41	0.682		-0.010
ST Attractiveness	-0.027	0.009	-3.15	0.002	-0.016	0.013	-1.34	0.178	-0.014	0.011	-1.01	0.218		-0.018
ST – LT Attractiveness	-0.018	0.008	-2.13	0.033	-0.015	0.012	-1.22	0.221	-0.022	0.011	-1.99	0.047		-0.025
ST Attractiveness w/ LT controlled	-0.029	0.009	-3.12	0.002	-0.019	0.013	-1.46	0.146	-0.020	0.012	-1.62	0.105		-0.025
Partner Attractiveness vs. Own	-0.023	0.008	-2.76	0.006	-0.015	0.012	-1.19	0.234	-0.024	0.012	-1.79	0.073		-0.016
Overall Aggregate of Partner Attractiveness														
Partner Attract w/ LT controlled	-0.031	0.009	-3.56	<.001	-0.017	0.013	-1.31	0.191	-0.024	0.012	-2.00	0.045		-0.023
Mean	-0.025				-0.016				-0.018					-0.020
Sample of 123	Continuous				Narrow Windows				Broad Windows					Robustness[1]
	γ	SE	*t*	*p*	γ	SE	*t*	*p*	γ	SE	*t*	*p*		γ
Physical Attractiveness	-0.017	0.009	-1.84	0.066	-0.014	0.013	-1.11	0.266	-0.006	0.012	-0.54	0.589		-0.010
ST Attractiveness	-0.023	0.009	-2.46	0.014	-0.019	0.013	-1.46	0.144	-0.016	0.012	-1.33	0.182		-0.018
ST – LT Attractiveness	-0.018	0.009	-2.01	0.045	-0.024	0.012	-1.97	0.049	-0.031	0.011	-2.65	0.008		-0.025
ST Attractiveness w/ LT controlled	-0.025	0.010	-2.56	0.010	-0.024	0.014	-1.77	0.076	-0.024	0.013	-1.90	0.057		-0.025
Partner Attractiveness vs. Own	-0.022	0.009	-2.56	0.011	-0.017	0.013	-1.28	0.201	-0.021	0.012	-1.74	0.083		-0.016
Overall Aggregate of Partner Attractiveness														
Partner Attract w/ LT controlled	-0.028	0.009	-3.16	0.002	-0.022	0.013	-1.66	0.097	-0.028	0.012	-2.26	0.024		-0.023
Mean	-0.022				-0.020				-0.021					-0.020

Notes. All $p < .01$ bolded and italicized; $p < .05$ bolded; $p < .10$ italicized. Models do not include random slopes, following Arslan et al.'s preregistration. See SOM R analyses for full model results and analyses with random slopes. 95% CI = γ ± 1.96 × SE

[1] These values from the robustness sample are from Table 1, main text, included here for purposes of comparison.

Figure S5.1 displays point estimates and confidence intervals when the 5-component measure of partner attractiveness is entered as a moderator, with partner LT attractiveness controlled. Point estimates across analyses are similar. But confidence interval widths predictably differ. Within the robustness sample, it is naturally smaller. In other analyses, confidence intervals are wider and, when random effects for fertility are also included, some include zero. But in no sense do the different confidence intervals imply fundamentally different conclusions.

Figure S5.1. Point estimate of effect size (circle) and confidence intervals (represented by lines) for the interaction between the 5-component aggregate measure of partner attractiveness and fertility, plotted for the robustness sample (continuous measure), the preregistered sample of 123 using the continuous measure, and the preregistered sample of 123 using narrow and broad fertile windows. Y-axis is effect size, where all predictors and outcomes are z-scored amd, hence, effect sizes can be meaningfully compared. LT mate attractiveness controlled in these analyses.



This situation is near-identical to the primary published example that Amrhein et al. used to illustrate their point about deriving false inferences from binary statistical significance. In that case, researchers concluded that, because they detected no significant association between use of anti-inflammatory drugs and new-onset atrial fibrillation, their results contrast with an earlier finding. In fact, point estimates were nearly identical; the prior study had greater sample size and, hence, a narrower confidence interval, one excluding zero. As Amrhein et al. noted, "it is … absurd to claim these [new] results were in contrast with the earlier results *showing an identical observed effect*" (p. 306; emphasis added). Similarly, it makes little sense to think that current analyses on a smaller sample are inconsistent with those on a larger sample, when effect size estimates across analyses agree.

Validity of fertility measures

Comparison of effects across analyses presented in Table S5 and Figure S5 are most meaningful if the fertility measures used in these analyses have similar validity. In fact, if all cycle days are used, the continuous measure has greater validity (Gangestad et al., 2016). But Arslan et al.'s fertility window measures eliminated days just before and after the estimated fertile phase (days on the "shoulders" of the fertile windows), which increases the validity of narrow and broad windows, relative to inclusion of those days. (The reason is simply because exclusion leaves in days with less ambiguous fertility status.) By contrast, analyses using a continuous measure retained all days. To estimate the validity of fertility measures in Arslan et al.'s sample, we used Gangestad et al.'s (2016) simulation sample used to compare validity of different measures. The sample we used assumes a rate of 8% anovulatory cycles (see Gangestad et al., 2016, for details). We used estimated probability of actually falling within a fertile window as a criterion of "true fertility" for estimates of validity coefficients. (Probability of conception given unprotected sex as a criterion produced near-identical values.):

- When the day of next menstruation is known (as in Arslan et al.'s sample of 143), Arslan et al.'s narrow and broad windows are estimated to have a mean validity coefficient of **.65** (.72 for narrow windows, .59 for broad windows).
- When the day of next menstruation is known, the continuous measure (based on backward counting) has an estimated validity of **.66.**
- Within Arslan et al.'s robustness sample, in which 67% of days have a known onset of next menstruation, and 33% have an estimated onset of next menstruation based on typical cycle length (their Table 3), the continuous measure has an estimated validity of **.62** (.66 for days with known onset, .54 for estimated onset). (This method assumes a correlation between self-reported typical cycle length and actual cycle length of .5 (see Gangestad et al., 2016).

These values assume that reports of onset of next menstruation are perfectly error-free. If reporting errors are minimal, estimated validity of these measures is roughly .6. Validity within Arslan et al.'s sample of 143 (and the preregistered subset of 123) is estimated to be slightly greater than validity within the robustness sample, but the difference is small.

S6. Estimation of effect sizes for women with relatively unattractive and attractive partners in Table 4: Detailed description

The estimated effects of fertility, expressed in *d*, for women with relatively unattractive and relatively attractive partners, were calculated as follows. We begin with calculations controlling for male partners' LT attractiveness:

The overall fertility effect is .0337. The partner sexual attractiveness × fertility effect is -.0235 (SE = .0079). Hence, fertility effects -1 *sd* and +1 *sd* from the mean of attractiveness are .0337+.0235 = .0572 and .0337 - .0235 = .0102. The within-woman standard deviation of EP sexual interests is .56. The standardized estimated within-woman effects, which also approximate the partial correlation between fertility and within-woman EP sexual interests (as other predictors account for minimal amounts of within-woman variance in EP interests), are .0572/.56 = .102 and .0102/.56 = .018. Disattenuation for imperfect validity of fertility measurement (validity = .6; Gangestad et al., 2016) and reliability of measurement of EP interests (reliability = .6; Arslan et al., 2018) yields values of .102/(.6 × √.6) = .220 and .018/(.6 × √.6) = .039. If we assume a 6-day fertile window and, on average, 23 days outside of the fertile window, these correlations translate to Cohen's *d* of .530 and .096.

To obtain estimates of the upper bound of a 95% confidence interval, we multiplied 1.96 times the SE of partner sexual attractiveness × fertility effect size and added this value to partner sexual attractiveness × fertility effect size (.0235 + 1.96 × .0079 = .0390). We then followed the steps described above to estimate effect sizes in *d* for women with partner sexual attractiveness -1 *sd* and +1 *sd* from the mean.

To obtain estimates of the lower bound of a 95% confidence interval, we multiplied 1.96 times the SE of partner sexual attractiveness × fertility effect size and subtracted this value to partner sexual attractiveness × fertility effect size (.0235 - 1.96 × .0079 = .0081). We then followed the steps described above to estimate effect sizes in *d* for women with partner sexual attractiveness -1 *sd* and +1 *sd* from the mean.

Values can be further disattenuated for imperfect reliability of partner sexual attractiveness, estimated to be .65 for the 5-component composite.

The same procedures were used to calculate effect sizes based on results from analyses in which male partners' LT attractiveness was not controlled. In those analyses, the overall fertility effect is .0340. The partner sexual attractiveness × fertility effect is -.0183 (SE = .0075).

S7. Effect size estimation based on models with no random slopes

	Control for LT attractiveness	
	Controlled	**Not controlled**
Point estimates		
No disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.51	0.45
1 *sd* above mean	0.09	0.16
Disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.56	0.49
1 *sd* above mean	0.03	0.13
Strongest effect within 95% CI		
No disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.62	0.55
1 *sd* above mean	-0.03	0.05
Disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.69	0.61
1 *sd* above mean	-0.11	-0.01
Weakest effect within 95% CI		
No disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.39	0.35
1 *sd* above mean	0.24	0.27
Disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.41	0.36
1 *sd* above mean	0.22	0.26

Notes. All estimates derived from mixed model analysis based on Arslan et al.'s procedures, in the robustness sample, using the 5-component measure of partner sexual attractiveness we created. No random slope for fertility modeled.

S8. Effect size estimation based on models using ST attractiveness as a moderator

	Control for LT attractiveness	
	Controlled	**Not controlled**
Point estimates		
No disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.55	0.48
1 *sd* above mean	0.08	0.15
Disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.60	0.52
1 *sd* above mean	0.02	0.11
Strongest effect within 95% CI		
No disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.69	0.62
1 *sd* above mean	-0.08	0.00
Disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.77	0.69
1 *sd* above mean	-0.18	-0.08
Weakest effect within 95% CI		
No disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.40	0.35
1 *sd* above mean	0.23	0.29
Disattenuation for imperfect reliability of partner attractiveness		
1 *sd* below mean	0.44	0.35
1 *sd* above mean	0.19	0.28

Notes. All estimates derived from mixed model analysis based on Arslan et al.'s procedures, in the robustness sample, using our recalculated measure of partner ST attractiveness measure. Random slope for fertility modeled.

S9. Simple effects and confidence intervals

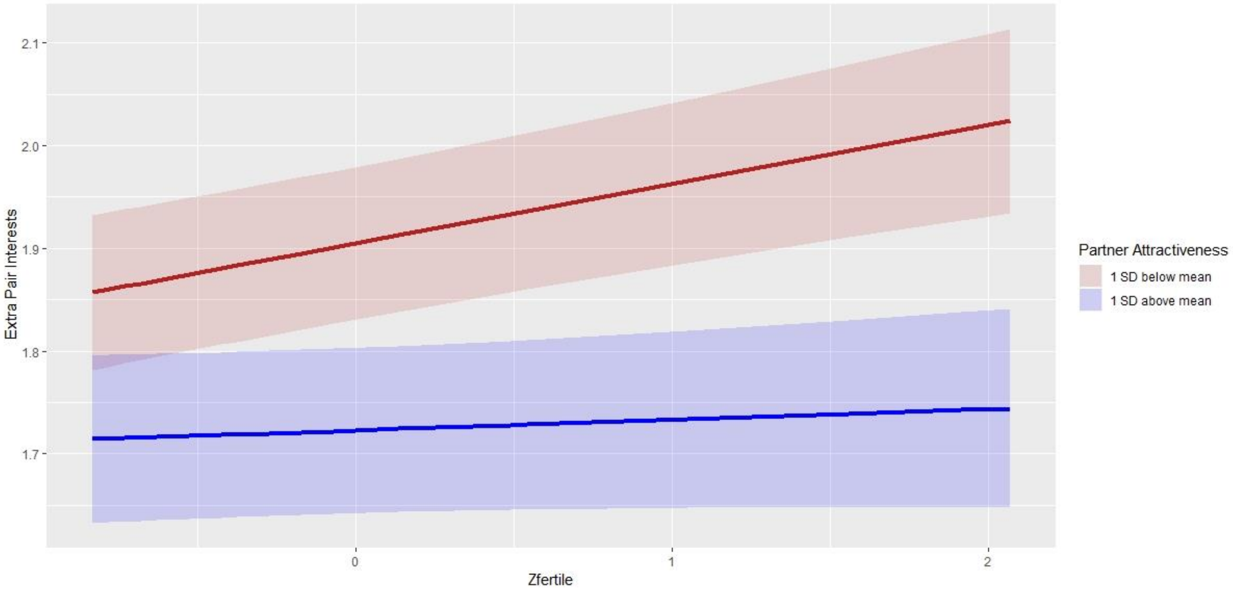
In mixed model analyses on the robustness sample (*N* = 429 naturally cycling women) with random slopes for fertility, we estimated simple effects for fertility on EP sexual interests for women with partners reported to be 1 *sd* above the mean and 1 *sd* below the mean on attractiveness. Results are given in the table below. As can be seen, fertility effects are estimated to be much larger for women with partners reported to be relatively unattractive than for women with partnered reported to be relatively attractive.

Table S9. Effect sizes, test statistics, and p-values for associations of fertility status with female extra-pair sexual interests within the robustness sample: Simple effects for women with partners 1 sd above and below the mean on the 5-component composite measure of partner attractiveness

Moderator	**1 *sd* above mean attractiveness**					**1 *sd* below mean attractiveness**			
	γ	**95% CI**	***t***	***p***		***γ***	**95% CI**	***t***	***p***
Physical Attractiveness	0.024	.000 to .047	1.99	0.047		0.046	.024 to .068	4.17	<.001
ST Attractiveness	0.016	-.007 to .039	1.33	0.183		0.052	.030 to .074	4.74	<.001
ST – LT Attractiveness	0.009	-.015 to .033	0.74	0.459		0.060	.037 to .082	5.21	<.001
ST Attractiveness w/ LT controlled	0.008	-.016 to .032	0.65	0.516		0.059	.037 to .082	5.18	<.001
Partner Attractiveness vs. Own	0.019	-.004 to .042	1.62	0.105		0.049	.028 to .071	4.51	<.001
Overall Aggregate of Partner Attractiveness									
Partner Attract w LT controlled	0.010	-.014 to .034	0.84	0.401		0.057	.035 to .079	5.13	<.001

Figure S9.1

Plot of association between *z*-scored fertility (x-axis) and extra-pair sexual interests (y-axis) for women with partners 1 *sd* below the mean on attractiveness (5-component composite measure) (red line) and for women with partners 1 *sd* above the mean on attractiveness (blue line). Shaded areas are 95% confidence intervals. Long-term attractiveness is controlled. As can be seen, the fertility effect on EP interests is driven by women with partners they report to be relatively unattractive.


2.1
2.0
1.9
1.8
1.7
Extra Pair Interests
0
1
2
Zfertile
Partner Attractiveness
1 SD below mean
1 SD above mean

S10. Effect size estimation: Prior studies vs. Arslan et al.

How do moderation effect sizes on EP sexual interests in Arslan et al.'s data set compare with effects observed in prior studies? This question is challenging to answer because of methodological differences across studies. Most past studies have used urinary luteinizing hormone (LH) surges to identify the fertile phase. In these studies, women for whom no LH surge is detected are excluded from the sample. The primary dependent variable used in these studies is a difference in reports of EP sexual interest over the past two days during "high fertility" (peri-ovulatory) and "low fertility" (luteal phase) sessions. In all studies, women have reported to the lab for initial sessions, and all but one (Haselton & Gangestad, 2006) had women return to the lab to complete all subsequent test sessions. By contrast, Arslan et al. used a counting method to measure fertility, had women provide daily reports of EP sexual interest, and recruited an online sample of women who never reported to a lab.

Prior published studies

Larson et al. (2012) provide a table with all past studies examining moderation of cycle shifts in EP sexual interest by partner features. Four studies explicitly examined moderation by partner attractiveness: Haselton & Gangestad (2006), Pillsworth & Haselton (2006), Gangestad et al. (2010), and Larson et al. (2012). All studies aside from Gangestad et al. (2010) had women themselves rate their partners' attractiveness. Gangestad et al. (2010) used 3rd party ratings of photographs. Larson et al. (2012) used both methods, but we used ratings by women in this study, as it best matches Arslan et al.'s procedures. (Including results from this study using 3rd party ratings yields a near-identical estimate of effect size.) Since Larson et al. (2012), Shimoda et al. (2018) examined moderation of EP interest by partner sexual vs. long-term attractiveness. We express effect size as the correlation between partner attractiveness with the high fertility – low fertility difference in EP interest, computed from test statistics.

	t-value	N	est r
Pillsworth & Haselton 2006	2.35	43	-0.37
Gangestad et al 2010	1.12	63	-0.14
Haselton & Gangestad 2006	3.08	24	-0.66
Larson et al. 2012	1.30	41	-0.21
Shimoda et al. 2018	2.32	35	-0.40
Average (weighted by square root of N)			-0.29

The weighted mean correlation is -.29. The standard deviation across Fisher r_z values is .19. Given sample sizes, the expected standard deviation due to sampling variability alone is .17; hence, we detect no meaningful heterogeneity across studies (though, given the small number of estimates, power to detect heterogeneity is very low.)

Larson et al. (2012) list two additional studies examining moderation by partner features, one examining moderation by partner developmental stability (Gangestad et al., 2005) and one examining moderation by partner facial masculinity (Gangestad et al., 2010). Inclusion of these two studies does not meaningfully alter the mean estimated effect size (-.30).

Arslan et al.

To estimate effects in Arslan et al.'s study, for each measure of partner attractiveness, we ran two models; first, a model with random slopes for fertility including the fixed partner attractiveness × fertility moderation effect; second, a model with random slopes for fertility not including this moderation effect. In the latter model, one expects that the random variation in fertility effects across women is increased to the extent that partner attractiveness does moderate the impact of fertility. Proportion of variance in fertility effects accounted for partner attractiveness is one minus the ratio of the fertility random effects variance estimates (τ_{11}) for the models including and excluding the fixed partner attractiveness × fertility interaction. (This random variance for the model including the fixed interaction is the estimated residual variance in fertility effects, including error. The random variance for the model excluding the fixed interaction includes variation driven by partner attractiveness; see, e.g. King et al., 2018). The correlation between partner attractiveness and fertility effects on EP interests, then, is estimated by the square root of this estimated variance accounted for. We used the robustness sample to estimate these correlations. Results are given below.

	Var ratio[a]	Est r
5-component composite	0.943	-0.24
Physical Attractiveness	0.995	-0.07
Short-term Attractiveness	0.963	-0.19
Short-term - Long-term attractiveness	0.940	-0.25
Short-term controlling for Long-term Attractiveness	0.939	-0.25
Partner Attractiveness vs. Self-attractiveness	0.972	-0.17
Average		-0.194

[a] Ratio of random slope variation with and without fixed effect for moderation effect modeled

As can be seen, the mean estimated correlation is -.19. The estimated correlation for arguably the most valid measure of partner attractiveness—the 5-component composite measure—is -.24.

Potential reasons for discrepancy

Once again, methods across studies varied. One key difference concerns measurement of fertility. Prior studies used methods estimated to have weighted mean validity of .85. Arslan et al. used a method estimated to have a validity of ~.6 (see SOM S5, Validity of fertility measurement). The correlations above disattenuated for imprecise measurement of validity are -.34 (past studies) and -.32 (mean across Arslan et al.'s measures) to -.40 (composite measure in Arslan et al.), respectively. These correlations are very similar.

Partner attractiveness and EP sexual interests are also measured imprecisely. In these regards, it is not obvious that Arslan et al.'s measures are any less valid than those in past studies. EP sexual interest was estimated by Arslan et al. to have a within-woman, cross-days reliability of .60. Reliability could have potentially been somewhat compromised by the online nature of data collection. At the same time, it is not clear that reliability of measurement in prior studies was higher. Disattenuation of the above correlations assuming a reliability of measurement of EP interests of .6 results in values of -.44 (past studies), -.42 (mean across Arslan et al.'s measures), and -.51

(composite of Arslan et al.'s measures). Further disattenuation for imprecise measurement of partner attractiveness assuming a validity of .8 yields values of -.55 (past studies), -.52 (mean across Arslan et al.'s measures), and -.64 (composite of Arslan et al.'s measures).

Arslan et al. obtained multiple daily reports per woman. Most previous studies have obtained two reports, albeit each asking about a two-day period. For this reason, Arslan et al.'s procedures could better estimate systematic variation in fertility effects. At the same time, other studies typically used in-lab questionnaires, as opposed to online questionnaires. The relative validity of the studies in this regard is unclear.

Naturally, disattenuated correlations have large standard errors. That said, imprecise measurement is known to attenuate manifest correlations, which accordingly are expected to underestimate correlations between perfectly measured variables. Estimates of those correlations, based on estimates of validity of measurement, are consistent with there being meaningful effects, both in prior studies and in Arslan et al.

Publication bias and selective reporting

Finally, of course, effect sizes in prior studies may well be overestimated due to publication bias and selective reporting (e.g., reporting of some measures within studies and not other measures). After differences in validity of fertility measurement are accounted for, effects estimated from prior studies and those estimated from Arslan et al.'s data are very similar. We nonetheless do not doubt that effects in prior studies are likely overestimates of true effects. For this reason, Arslan et al.'s data may offer the best estimates of true effect size to date. Once again, the point estimates derived from those data are substantial and meaningful while, at the same time, bracketed by wide confidence intervals.

S11. Mate retention tactics

Arslan et al. found little over-time reliability in their measure of male mate retention tactics. One reason may be that they included items tapping proprietariness and attentiveness in the same index. As noted earlier, previous studies have consistently separated two minimally covarying broad dimensions of mate retention tactics, proprietariness and attentiveness. Gangestad et al. (2014) found that only a factor of male proprietariness increased with women's extra-pair sexual interests as a function of fertility status. Arslan et al. explicitly acknowledged Gangestad et al.'s (2014) findings and, accordingly, ran exploratory analyses on individual items: "[W]e additionally calculated all analyses by item in a purely exploratory manner. 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" (p. xxx). In our view, Arslan et al. should have run analyses on separate components of mate retention, in accord with previous research appearing before their preregistration (e.g., Gangestad et al., 2002; Haselton & Gangestad, 2006), and considered the results of their own exploratory analyses before concluding that they could not replicate previous research findings.

Arslan et al. did not report the results of their exploratory analyses. We created two simple composites from the four mate retention items: two items were averaged to create a proprietariness measure ("My partner asked me with whom I spent the day"; "My partner was jealous of my contact with other men"); two other items were averaged to create a measure of attentiveness (e.g., "My partner told me he loved me"; "My partner showed me that he was sexually attracted to me"). A principal components analysis on the within-woman cross-time variation in the four items yielded two clearly separable components, with appreciable loadings by two items each. See below.

We then used the same code to examine predictors of EP sexual interests to examine moderation effects on male proprietariness. All measures of male sexual attractiveness significantly moderated the impact of fertility status on male proprietariness. Effects remain robust when random slope variation for fertility is modeled.

Table S11a reports effects using our primary corrected measure of partner ST attractiveness was used. Table S11b reports effects using Arslan et al.'s measure of partner ST attractiveness, with participants not responding to the item concerning sexual satisfaction removed (see SOM S4).

Table S11a. Test statistics and p-values for moderation of associations of fertility status with male propriatariness by partner attractiveness, robustness sample

Moderator	**No random slopes**					**Fertility random slope**			
	γ	**95% CI**	***t***	***p***		***γ***	**95% CI**	***t***	***p***
Physical Attractiveness	-0.046	-.080 to -.012	-2.65	0.008		-0.052	-.095 to -.009	-2.39	0.017
ST Attractiveness	-0.044	-.077 to -.010	-2.52	0.012		-0.046	-.089 to -.003	-2.11	0.035
ST – LT Attractiveness	-0.053	-.089 to -.017	-2.88	0.004		-0.051	-.097 to -.005	-2.19	0.029
ST Attractiveness w/ LT controlled	-0.057	-.094 to -.020	-3.03	0.002		-0.058	-.105 to -.005	-2.44	0.015
Partner Attractiveness vs. Own	-0.077	-.110 to -.044	-4.52	<0.001		-0.077	-.119 to -.035	-3.62	<0.001
Overall Aggregate of Partner Attractiveness									
Partner Attract w LT controlled	-0.086	-.121 to -.051	-4.77	<0.001		-0.087	-.131 to -.043	-3.83	<0.001
Partner Attract w/o LT controlled	-0.074	-.107 to -.040	-4.32	<0.001		-0.075	-.117 to -.033	-3.51	<0.001

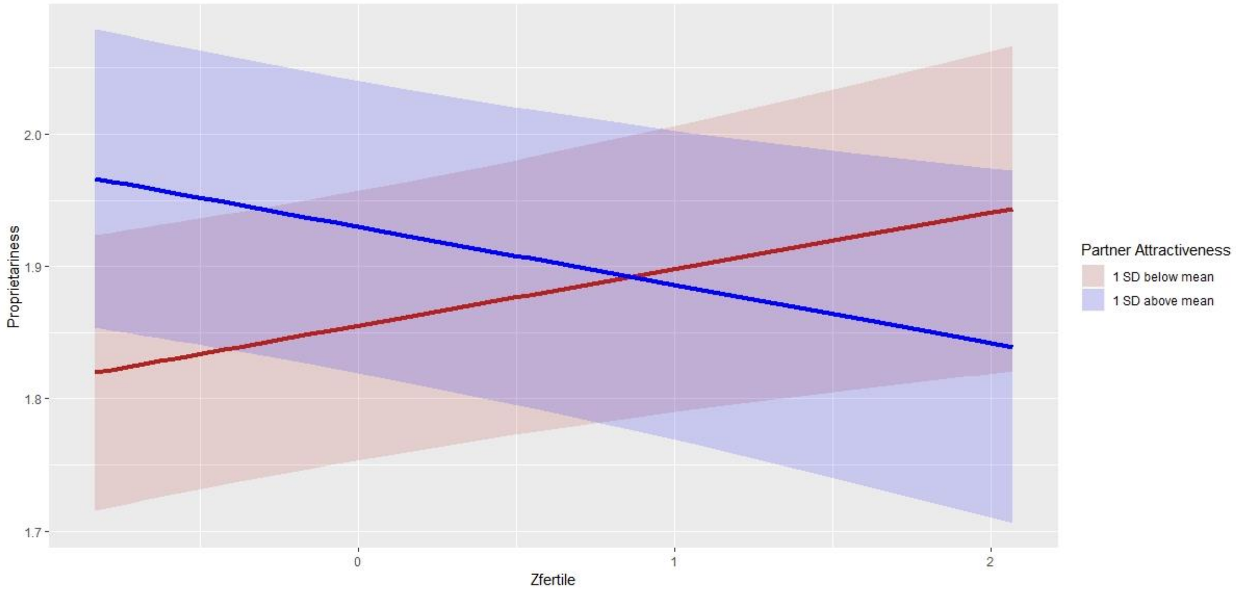
ST = short-term; LT = long-term attractiveness. *N* of normally cycling women = 429.
Mean of z-scored components used as measure of ST attractiveness. See SOM S4. Full results of all analyses provided in SOM. Effect size estimates are taken from analyses using *z*-scored predictors.

Table S11b. Test statistics and p-values for moderation of associations of fertility status with male propriatariness by partner attractiveness, robustness sample

Moderator	**No random slopes**					**Fertility random slope**			
	γ	**95% CI**	***t***	***p***		***γ***	**95% CI**	***t***	***p***
ST Attractiveness	-0.057	-.092 to -.022	-3.17	0.002		-0.062	-.107 to -.017	-2.70	0.007
ST – LT Attractiveness	-0.063	-.100 to -.026	-3.30	0.001		-0.061	-.108 to -.014	-2.52	0.012
ST Attractiveness w/ LT controlled	-0.071	-.109 to -.033	-3.64	<.001		-0.074	-.122 to -.025	-2.98	0.003

ST = short-term; LT = long-term attractiveness. *N* of normally cycling women = 429.
Arslan et al. sum of components used as measure of ST attractiveness, removing participants who did not respond to sexual satisfaction item. See SOM S4. Full results of all analyses provided in SOM. Effect size estimates are taken from analyses using *z*-scored predictors.

Figure S11.1. Plot of association between *z*-scored fertility (x-axis) and proprietariness (y-axis) for women with partners 1s below the mean on attractiveness (5-component composite measure) (red line) and for women with partners 1s above the mean on attractiveness (blue line). Shaded areas are 95% confidence intervals. Long-term attractiveness controlled.



Attentiveness

No moderation effects were found with attentiveness. This pattern of effects—moderation on proprietarieness but none on attentiveness—is consistent with findings of Gangestad et al. (2014) and Haselton and Gangestad (2006). Table S11c reports effects using our primary corrected measure of partner ST attractiveness was used. Effects using Arslan et al.'s measure of partner ST attractiveness, with women who did not respond to the item concerning sexual satisfaction removed (see SOM S4), yielded similar results.

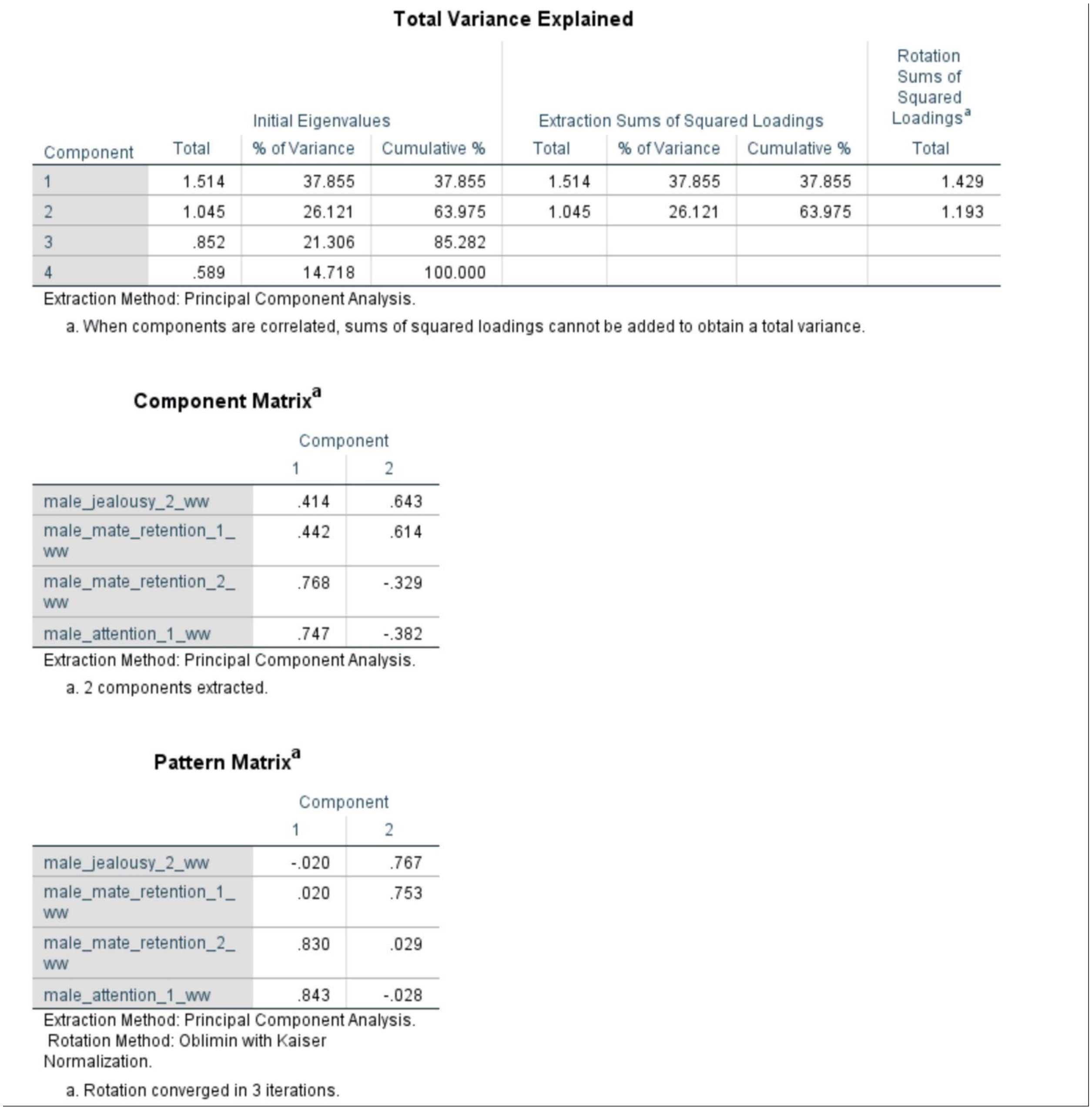
Table S11c. Test statistics and p-values for moderation of associations of fertility status with male attentiveness by partner attractiveness, robustness sample

Moderator	**No random slopes**					**Fertility random slope**			
	γ	**95% CI**	***t***	***p***		***γ***	**95% CI**	***t***	***p***
Physical Attractiveness	0.015	-.035 to .065	0.60	0.549		0.012	-.050 to .074	0.38	0.707
ST Attractiveness	0.014	.036 to .064	0.56	0.573		0.014	-.049 to .077	0.44	0.659
ST – LT Attractiveness	0.022	-.032 to .075	0.79	0.430		0.019	-.048 to .086	0.56	0.577
ST Attractiveness w/ LT controlled	0.020	-.034 to .075	0.73	0.463		0.019	-.050 to .088	0.54	0.586
Partner Attractiveness vs. Own	-0.027	-.076 to .023	-1.06	0.289		-0.024	-.086 to .038	-0.76	0.448
Overall Aggregate of Partner Attractiveness									
Partner Attract w LT controlled	0.008	-.044 to .060	0.30	0.762		0.007	-.059 to .072	0.20	0.837

Principal components analysis of 4 male mate retention items

Male_jealousy_2_ww: "My partner was jealous of my contact with other men."
Male_mate_retention_1_ww: "My partner asked me with whom I spent the day."
Male_mate_retention_2_ww: "My partner told me he loved me."
Male_attention_1_ww: "My partner showed me that he was sexually attracted to me."

All items are scaled as responses relative to a woman's overall mean response to the item (i.e., as "within-woman" variations.



Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	1.514	37.855	37.855	1.514	37.855	37.855	1.429
2	1.045	26.121	63.975	1.045	26.121	63.975	1.193
3	.852	21.306	85.282				
4	.589	14.718	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

Component Matrix[a]

	Component	
	1	2
male_jealousy_2_ww	.414	.643
male_mate_retention_1_ww	.442	.614
male_mate_retention_2_ww	.768	-.329
male_attention_1_ww	.747	-.382

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Pattern Matrix[a]

	Component	
	1	2
male_jealousy_2_ww	-.020	.767
male_mate_retention_1_ww	.020	.753
male_mate_retention_2_ww	.830	.029
male_attention_1_ww	.843	-.028

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 3 iterations.

References

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328. https://doi: 10.3389/fpsyg.2013.00328

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*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

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*, 302-308. https://doi.org/10.1016/j.evolhumbehav.2014.02.012

Gangestad, S. W., Haselton, M. G., Welling, L. L. M., Gildersleeve, K., Pillsworth, E. G., Burriss, R. L., 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*, *37*, 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 of London* B, *269*, 975-982. https://doi.org/10.1098/rspb.2001.1952

Gangestad, S. W., Thornhill, R., & Garver-Apgar, C. E. (2005). Women's sexual interests across the ovulatory cycle depend on primary partner developmental instability. *Proceedings of the Royal Society of London* B, *272*, 2023-2027. https://doi.org/10.1098/rspb.2005.3112

Gangestad, S. W., Thornhill, R., & Garver-Apgar, C. E. (2010). Men's facial masculinity predicts changes in their female partners' sexual interests across the cycle, whereas men's intelligence does not. *Evolution and Human Behavior*, *31*, 412-424. https://doi.org/10.1016/j.evolhumbehav.2010.06.001

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*, 509-518. https://doi.org/10.1016/j.yhbeh.2005.10.006

King, K., Kim, C. S., McCabe C., & Lane, S. P. (2018) Stepwise methods can limit power for hypothesis tests of cross-level interactions. *Psyarxiv*. https://psyarxiv.com/92btp/

Larson, C. M., Pillsworth, E. G., & Haselton, G. M., (2012). Ovulatory shifts in women's attractions to primary partners and other men: Further evidence of the importance of primary partner sexual attractiveness. *PLoS ONE*, *7*, e44456 https://doi.org/10.1371/journal.pone.0044456

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*, 247–258. https://doi.org/10.1016/j.evolhumbehav.2005.10.002

Shimoda, R., Campbell, A., & Barton, R. A. (2018). Women's emotional and sexual attraction to men across the menstrual cycle. *Behavioral Ecology*, *29*, 51-59. https://doi.org/10.1093/beheco/arx124