Using 26 thousand diary entries to show ovulatory changes in sexual desire and behaviour

forthcoming in Journal of Personality and Social Psychology: Personality Processes and Individual Differences

Authors: Ruben C. Arslan[1,2,3,*], Katharina M. Schilling[2], Tanja M. Gerlach[2,3†], Lars Penke[2,3†]
* corresponding author
† LP and TMG share the last authorship.

Affiliations:

1. Center for Adaptive Rationality
   Max Planck Institute for Human Development
   14195 Berlin, Germany
2. Biological Personality Psychology,
   Georg Elias Müller Institute of Psychology,
   University of Goettingen,
   37073 Göttingen, Germany
3. Leibniz ScienceCampus Primate Cognition,
   37073 Göttingen, Germany

Author Note: This research has been previously presented at the Human Behavior and Evolution Conference (2017), at the European Human Behaviour and Evolution Association (2016), and the Congress of the German Society of Psychology (2016). Preliminary results from this study formed the basis of KMS' master thesis. The authors have no conflicts of interest.

Acknowledgements: We thank Hanne Straus for her help collecting the data on hormonal contraception, Silvia Bradatsch for extracting the sample sizes from the meta-analysis by Gildersleeve et al. (2014a), and Aileen Marske for helping with the supportive materials. We thank Paul-Christian Bürkner, author of brms, and Ben Bolker, co-author of lme4, for their free statistical packages and advice on using them. We thank Isabelle Habedank, Maren Fußwinkel, Sarah J. Lennartz, Steven Gangestad, Ben Jones, and two anonymous reviewers for their helpful comments on earlier versions of this manuscript.

Contact information:
Ruben Arslan
Center for Adaptive Rationality
Max Planck Institute for Human Development
Lentzeallee 94
14195 Berlin
Email: **ruben.arslan@gmail.com**

Study documentation: https://osf.io/kd26j, doi:10.17605/osf.io/kd26j
Files to reproduce the questionnaires (original and English translations) and study structure, the preregistration, and synthetic data.

Supportive website: https://rubenarslan.github.io/ovulatory_shifts, doi:10.5281/zenodo.1243038
Reproducible documentation of all data cleaning procedures, all reported analyses, and several supportive analyses.

Direct links to materials: https://osf.io/pbef2, doi:10.17605/osf.io/kd26j
A link list to all materials mentioned in the manuscript.

Abstract

Previous research reported ovulatory changes in women's appearance, mate preferences, extra- and in-pair sexual desire and behaviour, but has been criticised for small sample sizes, inappropriate designs, and undisclosed flexibility in analyses. In the present study, we sought to address these criticisms by preregistering our hypotheses and analysis plan and by collecting a large diary sample. We gathered over 26 thousand usable online self-reports in a diary format from 1043 women, of which 421 were naturally cycling. We inferred the fertile period from menstrual onset reports. We used hormonal contraceptive users as a quasi-control group, as they experience menstruation, but not ovulation. We probed our results for robustness to different approaches (including different fertility estimates, different exclusion criteria, adjusting for potential confounds, moderation by methodological factors). We found robust evidence supporting previously reported ovulatory increases in extra-pair desire and behaviour, in-pair desire, and self-perceived desirability, as well as no unexpected associations. Yet, we did not find predicted effects on partner mate retention behaviour, clothing choices, or narcissism. Contrary to some of the earlier literature, partners' sexual attractiveness did not moderate the cycle shifts. Taken together, the replicability of the existing literature on ovulatory changes was mixed. We conclude with simulation-based recommendations for reading the past literature and for designing future large-scale preregistered within-subject studies to understand ovulatory cycle changes and the effects of hormonal contraception. Interindividual differences in the size of ovulatory changes emerge as an important area for further study.

Keywords: ovulatory cycle shifts, sexual desire, diary study, hormonal contraception, evolutionary psychology

Introduction

Theoretical Background

Personality, behaviour, sexual desire, attractiveness, mate preferences and mate choices vary between and within persons (Fleeson, 2001, 2004; Gerlach, Arslan, Schultze, Reinhard, & Penke, in press). While copious research has identified antecedents of interindividual variation (Zietsch, Lee, Sherlock, & Jern, 2015), intraindividual variation is still often viewed as mere chance fluctuation or response to situational demands. Systematic endogenous causes of intraindividual variation are worthy of further study.

In the evolutionary psychology literature, the menstrual cycle has been suggested as one such influence on psychological state fluctuations in women (Gangestad & Thornhill, 2008). Menstrual cycle changes in attractiveness, mate preferences, and sexual desire, as well as men's reactions to those changes have been interpreted as evidence for adaptations formed by sexual selection and sexually antagonistic coevolution, i.e. arms races between the sexes. However, to this day debate continues over the existence and extent of such changes (W. Wood, Kressel, Joshi, & Louie, 2014). In this paper, we have the twin goals of reviewing methodological problems with commonly used approaches and addressing them in a high-powered, preregistered replication study. Because our study was preregistered in March 2014, the introduction of this manuscript reflects our reading of the literature at that point in time. We review recent theoretical and empirical developments in the discussion.

Do human females show oestrus?

Human women do not develop garish sexual swellings or other prominent changes around ovulation, unlike their closest cousins, the chimpanzees (Deschner, Heistermann, Hodges, & Boesch, 2003). Moreover, human women and several other primates exhibit *extended sexuality,* that is they have sex outside the fertile window, not just during a period of oestrus or *heat* (Dixson, 2012). However, other, less conspicuous endocrine, behavioural, physiological and psychological changes happen over the course of the menstrual cycle and some peak when women are fertile (Gangestad & Simpson, 2000; Haselton & Gildersleeve, 2016). This led Gangestad and Thornhill (2008) to argue that the differentiation of functional and physiological aspects of fertile phase sexuality merits being called oestrus.

The good genes ovulatory shift hypothesis

The ovulatory shift hypothesis posits that women's mate preferences and choices vary with their fertility status. It is a central functional differentiation predicted under the human oestrus perspective (Gangestad & Thornhill, 2008). According to this theory, women would optimise their reproductive potential by choosing to be with partners who will invest in offspring during non-fertile times and choosing, if necessary, other, extra-pair males with *good genes* to provide their offspring's genes, i.e. to have sex with during the fertile phase. To differentiate this theoretically predicted *ovulatory shift in mate preferences to obtain good genes, potentially from extra-pair copulations* (Pillsworth & Haselton, 2006a) from simpler, generalized increases in sexual desire or *libido* in the fertile phase, we will call this theory *good genes ovulatory shift hypothesis* (GGOSH).

The theoretical concept of *good genes* is meant to index genetic qualities that women should want their offspring to inherit. The concept includes dyadic genetic fit (e.g., matching immunocompetence genes), genetic fit to the current environment, and few harmful mutations. It has no direct correspondence in the evolutionary genetic literature and some purported indicators of *good genes* are controversial (Arslan & Penke, 2015). Several male characteristics have been argued to indicate *good genes.* Cycle studies have then reported fertile phase increases in preferences for these traits, which include masculinity, low fluctuating asymmetry (Scheib, Gangestad, & Thornhill, 1999), and various measures of attractiveness (Gildersleeve, Haselton, & Fales, 2014a; Haselton & Gangestad, 2006; Larson, Haselton, Gildersleeve, & Pillsworth, 2013; Pillsworth & Haselton, 2006b). In laboratory studies, fertile phase shifts towards preferences for male stimuli with such characteristics (photos, videos, voice samples), have been cited as support for GGOSH (Gildersleeve et al., 2014a).

Rationale for the present study

In our study, we sought to replicate and extend previous results from field studies of naturally cycling women commonly cited as evidence of a differentiation of fertile phase sexuality. These field studies reported evidence for changes in female sexual interests and appearance across the cycle. Central results in these studies served as the rationale for the preregistration of our study, but not all our hypotheses are direct replications of previously significant effects in the literature. Rather, they reflect our understanding of the theoretical predictions made by the previous literature. Some predictions go beyond what was previously shown; we explicitly note this where applicable and return to this in the discussion.

Extra-pair desire and behaviour

Gangestad, Thornhill, and Garver (2002) asked 51 naturally cycling women (i.e., not using hormonal contraceptives) to report their sexual interests and fantasies once in the fertile and once in the non-fertile phase. Women reported substantially greater attraction to and fantasies about men other than primary partners when fertile.

In a sample of 54 couples and using the same study design, Gangestad, Thornhill, and Garver-Apgar, 2005 additionally reported support for a predicted moderator effect. Women showed stronger fertile phase increases in attraction to other men if paired with relatively asymmetrical primary partners. In a diary study, Haselton and Gangestad (2006) asked 38 naturally cycling women to provide daily reports of sexual interest and feelings for 35 days. Women reported that they were more attracted to and flirted more often with men other than primary partners on higher fertility days if their partner's sexual attractiveness was low.

In-pair desire and behaviour

According to the ovulatory shift hypothesis, women whose long-term partners display indicators of "good genes" do not benefit from engaging in what Pillsworth and Haselton (2006a) call a dual mating strategy. The authors predicted such women should instead experience ovulatory increases in in-pair desire. Findings were mixed, with some showing the predicted moderated shifts (Gangestad et al., 2005; Pillsworth, Haselton, & Buss, 2004) while others did not (Gangestad et al., 2002; Pillsworth & Haselton, 2006b). Gangestad et al. (2002) found that women did not experience significantly higher levels of overall sexual desire when fertile, but tended to initiate and have more sex with their partners as ovulation neared.

Male mate retention

Because female extra-pair sex might lead her primary partner to involuntarily invest parental care and resources into offspring sired by an extra-pair mate, counter-adaptations to the aforementioned shifts were predicted (Pillsworth & Haselton, 2006a). Gangestad et al. (2002) correspondingly reported that *prohibitive* (i.e. jealousy) and *persuasive* (i.e. affection) male partners' mate retention tactics increased during the fertile phase. Haselton and Gangestad (2006) replicated these results, whereas Pillsworth and Haselton (2006b) only reported fertile phase increases in persuasive tactics. These tactics were exhibited primarily by partners of women who perceived their partners to be low in sexual attractiveness relative to investment attractiveness.

Self-perceived desirability, clothing choices, and self-esteem

Although obvious outward signals of fertility are absent in humans, some studies report evidence of subtle ovulatory cues in human females and conclude that ovulation may not be perfectly concealed. Haselton and Gangestad (2006) reported that women perceived themselves to be more attractive when fertile. Haselton et al. (2007) further predicted and found fertile phase increases in grooming and attractive clothing choices in a sample of 30 partnered women who were photographed at high and low fertility. Schwarz and Hassebrauck (2008) replicated and extended this study. In a sample of 40 women who completed a daily questionnaire over 31 days, participants rated their perceived attractiveness, and their clothing style on the dimensions "figure-hugging", "sexy", and "permissive". They were also instructed to take one photo of themselves each day. Men then rated these photos for clothing style and physical attractiveness. Women perceived themselves and were perceived by men to be dressed more provocatively on their fertile days. In another replication, using 88 women tested twice, Durante, Li, and Haselton (2008) reported evidence that women prefer clothing that is more revealing and sexy during the fertile phase, as shown in full-body photographs and drawn illustrations of what they would wear to a hypothetical social event that evening. Interestingly, Hill and Durante (2009) reported in two samples of 52 and 59 women tested twice that self-esteem decreased around ovulation. They reported this to be related to the willingness to spend money on enhancing one's own attractiveness. However, changes in general self-esteem can also be taken as intraindividual variation in daily mood, which might occur as a non-adaptive side-effect of hormonal changes. We were thus interested to find out whether any changes in sexual desire, self-perceived desirability, and clothing choices were independent of or larger than any changes in self-esteem.

Intrasexual competitiveness

Durante et al. (2008) interpreted their results discussed above as evidence of increased intrasexual competitiveness, i.e. women altering their physical appearance to enhance their ability to compete with other women. We speculated that, if intrasexual competitiveness during the fertile phase were increased, we might also detect this in narcissistic personality states, as conceptualized in the two-dimensional narcissistic admiration and rivalry concept (NARC; Back et al., 2013). Narcissistic admiration is thought to be linked to the desire to attain social status, and evoke social interest. Narcissistic rivalry is thought to be linked to motivations to defend one's social status against others. In the context of our study, to test the prediction of increased intrasexual competitiveness in the fertile phase (Durante et al., 2008) in a novel way, we reformulated narcissistic state items for both NARC dimensions to refer exclusively to comparisons with other women instead of people in general.

Methodological issues

The psychological literature on ovulatory changes has been criticised and hotly debated. Two meta-analyses based on overlapping data both concluded that publication bias afflicts research on ovulatory shifts in mate preferences, as may be the case for most of the scientific literature (Fanelli, 2011; Ferguson & Brannick, 2012). However, one team of investigators (Gildersleeve et al., 2014a) concluded that all evidence taken together suggested robust shifts in mate preferences, even after including studies freed from the file drawer and adjusting for bias. Another team (W. Wood et al., 2014) concluded further bias and methodological artefacts implied that any non-negligible effects were, in fact, overestimated. Our study focuses on different outcomes than these meta-analyses, but many problems discussed therein pertain to the designs commonly used to study ovulatory change more generally, irrespective of specific outcomes and research questions. Thus, they also informed our approach. In the following, we summarise several methodological issues brought to the fore by this debate.

Researcher degrees of freedom can lead to false positives

Many psychological studies do not replicate in exact replications (Open Science Collaboration, 2015). Potential sources of bias are *researcher degrees of freedom* in specifying hypothesis, methodology, and statistical approach after seeing the data. Journals and researchers tend to preferentially publish and cite significant counter-intuitive results, leading to warped incentives (Simmons, Nelson, & Simonsohn, 2011).

Recent debate in the menstrual cycle literature has specifically highlighted flexibility in the definition of the fertile window, but more general problems such as reporting only variables showing significant associations and stopping data collection conditional on significance could also affect the literature. Surveys of psychological researchers show that some research practices now deemed questionable were widespread (John, Loewenstein, & Prelec, 2012) and meta-analyses show publication bias. Both sides in the ovulatory cycle debate acknowledge bias (Gangestad, 2016; Harris, Pashler, & Mickes, 2014; W. Wood et al., 2014) but do not agree on whether and how it can be adjusted for (Gildersleeve, Haselton, & Fales, 2014b; Harris et al., 2014) in order to obtain trustworthy bias-corrected estimates (Inzlicht, Gervais, & Berkman, 2015; van Elk et al., 2015). The debate surrounding this has at times turned vitriolic, because the often used term *p-hacking* has connotations of intentional mischief, but it is clear from simulations (Smaldino & McElreath, 2016) and intuition (Gelman & Loken, 2014) that flexibility will lead to bias even without ill intentions, as long as odds of publication and career success can hinge on whether results turn out statistically significant. Ultimately, although methods such as the *p-curve* (Gildersleeve et al., 2014b) can offer suggestive evidence of replicability, the true tests of replicability are *preregistered* replication studies in which

hypotheses, methods and statistical approach are fixed before the data are collected, preventing *researcher degrees of freedom* from skewing results.

Estimating the day of ovulation and the fertile window

There is wide variability in the approaches used to estimate women's fertile windows. Gildersleeve et al., (2014a) reviewed these approaches and problems associated with them. Gangestad et al. (2016) recommend that researchers abandon windows altogether and instead estimate continuous probabilities of being fertile. Flawed recall of the last menstrual onset, accuracy being as low as 57% (Wegienka & Baird, 2005), remains a problem. Moreover, menstrual cycle lengths vary within person, so that recalled average cycle length correlates only ~.5 with the length of individual cycles (Blake, Dixson, O'Dean, & Denson, 2016; Gangestad et al., 2016). Because of flawed recall and because the follicular phase leading up to ovulation is more variable than the luteal phase (Fehring, Schneider, & Raviele, 2006), the more convenient method (forward counting from the last menstrual onset) is also more imprecise (Gangestad et al., 2016). Backward counting to ovulation from the *observed* next menstrual onset should hence be more accurate, with a validity for estimated fertility as high as ~.7 (Gangestad et al., 2016). Blake et al. (2016) report much lower validities (~.2-.3), using luteinising hormone (LH) surges as the criterion in a small sample of ~100 women, but our re-analyses of their data (see supportive website, **osf.io/pbef2**; Arslan, 2018) using the hedged fertile window estimate, as in Gangestad et al. (2016) and our study, showed a validity (.57) consistent with Gangestad et al.'s (2016) estimates.

For researchers, backward counting has the added benefit that women who count days as part of their contraception regimen cannot do it prospectively, perhaps reducing awareness and thus demand characteristics. Still, counting-based estimates of conception probability derive from forward-counted actuarial values which are then reversed (Gangestad et al., 2016). Ideally, actuarial estimates would be backward-counted, too. Given these concerns and that even the ~.7 validities are not very high, should not all research switch to more direct measurements?

Transvaginal ultrasonography (e.g., Caruso et al., 2014) is the gold standard measure for ovulation. While used to pinpoint ovulation in the context of cycle anomalies and fertility treatment, the increased validity may not be worth the substantially increased costs of measurement in psychological studies on average cycle patterns. More commonly, test strips to assess ovulation via luteinising hormone (LH) surges in urine are employed. Usually, they are used within a stretch of days deemed most likely to contain the day of ovulation by forward counting. Women who do not experience a surge in that time can then be excluded. Although LH surges can be detected very reliably, there is variation in the timespan between the surge and ovulation. Therefore, Gangestad et al. (2016) roughly estimate this method has validities around .8-.9. Although more valid and

comparatively cheap, drawbacks of this approach include that participants need to use the LH strips, familiar to many women. This makes it difficult to keep the research question opaque, potentially increasing demand characteristics perceived by the participants.

It is also possible to measure estrogen and progesterone in blood, urine, or, - as most commonly done in psychology - saliva. These hormones are strong candidates for the mechanisms behind ovulatory change. To compare this method to counting methods, we cannot simply look to the validity of fertile phase assessments via hormone assays, because this is not how hormone assays are normally used. While the estrogen-to-progesterone ratio can predict the LH surge (Baird et al., 1995), researchers usually directly predict outcomes from hormone levels. Researchers commonly report the coefficient of variation (CV) of their hormonal assays, a standard measure of reliability in biochemistry. Intra- and inter-assay CVs varies across labs and assays (e.g. for estradiol ~7/7% in Jones et al., 2018, 8/11% in Grotzinger et al., 2017), but arguably these do not directly map to the reliability estimate of interest in the ovulatory shift literature, namely the reliability of the changes measured in these hormones. In our re-analyses of the OCMATE data (Jones et al., 2018; see supportive website, **osf.io/pbef2**), these reliabilities of change were .83 for estradiol and .86 for progesterone.

To summarise, direct hormonal measurements have superior reliability and allow researchers to understand the mechanisms underlying ovulatory changes. However, they bring a new set of complexities with them, tie the research to a lab, make the research question less opaque, decrease anonymity, and are costlier than counting methods. Counting methods in online designs can compete with hormonal assays in lab-based studies in terms of statistical power, if we consider that more women and days can be sampled this way. Ultimately, lower reliability using the counting method can be compensated by sufficiently increasing sample size, and underestimation of effect sizes can be statistically accounted for (Gangestad et al., 2016). However, potential sources of bias such as a correlation between anovulation and moderator variables could bias moderator tests (see below for an example). Unless such biases are revealed in future studies, all methods should lead to converging evidence.

Between-subject designs to study a within-subject process

Many past studies have used between-subject designs to study a within-subject process, ovulation (Gangestad et al., 2016). Even when sample sizes are large, selection bias could confound any identified effects. One possible scenario could be that a common cause, for instance genetic makeup or a disease, makes women anovulatory and lowers their sexual desire. This could lead researchers to mistake a between-subject difference for an ovulatory change. Another potential problem might be that increased social activity during the fertile phase (Haselton & Gangestad, 2006)

could make fertile women less likely to participate in a survey study, biasing the sample towards women who experience smaller changes. Further, cross-sectional designs can never reliably measure individual differences in the size of ovulatory changes. They may also lead to the use of outcome measures that measure a trait component, but not a state component, reliably. This can be avoided by using established measures tested on within-subject data. Indeed, many of the above problems are minor and could potentially be avoided or adjusted for, but given that within-subject studies do not have these problems and are no longer hard to implement, they seem the superior option. Most crucially, however, between-subject studies have far too low statistical power at typical samples sizes, as shown by Gangestad et al. (2016).

Lack of power or implausible effect size expectations

The average menstrual cycle study to date is underpowered to detect anything but very large changes (Gangestad et al., 2016). At the same time, most researchers seem to agree that ovulatory changes are, if anything, subtle. In this situation, many plausible and interesting effect sizes will be missed, and reported effects will tend to be overestimates. If we desire theoretical progress, we need to narrow down effect sizes to disambiguate between theories that predict no, minimal, small, medium, or large ovulatory changes in certain outcomes. Thus, the literature would benefit from narrower confidence intervals to resolve theoretical debates over evolutionary function. Even for larger effects, typical cycle studies are underpowered, because of the combination of suboptimal design aspects and small sample size (median $N = 48$ in Gildersleeve, Haselton, and Fales, 2014; mean $N = 49$ in the studies we sought to replicate). For between-subject studies planning to achieve 80% power to detect a Cohen's d of 0.4 with a backward-counted conception probability estimate, Gangestad et al. (2016) recommend a sample size of 1,143.

No differentiation of women by reproductive intentions and contraception method

W. Wood et al. (2014) pointed out that the most uniquely human aspect of menstrual cycles may be women's exertion of control over their cycle and fertility to adapt to cultural, societal, and their own needs. Although they provide no specific recommendations how this should change research practices, we note that most studies do not differentiate between naturally cycling women who use barrier methods, awareness-based methods, or simply no contraception. Among women who do not use contraception, there may be women who are actively trying to conceive and would usually be excluded, but also those who do not mind risking a conception. Most studies also do not report asking women whether they track their fertility or menstrual cycle by counting with an app or calendar in addition to a primary contraceptive. If women are aware of their fertility status, their answers in the fertile phase might differ spuriously due to changed behaviour (e.g. avoiding sex or using condoms, or

seeking sex to conceive), heightened self-awareness for sexual thoughts and fantasies, demand characteristics, or personal theories on how their menstrual cycle affects them.

Lack of control group

Unfortunately, many cycle studies exclude women using hormonal contraceptives (HC) from taking part or from analysis, even though they can serve as a quasi-control group that experiences menstruation but not ovulation and the concurrent hormonal changes. A quasi-control group is also useful as an empirical baseline for the false discovery rate: if researchers found as many 'ovulatory' changes among HC users as among naturally cycling women, this would serve as feedback that the analysis procedure might entail false positives or invalid conclusions about the hormonal processes driving the changes. Apart from being a helpful methodological feature, including HC users allows researchers to more directly test whether, say, shifts in mate preferences or extra-pair desire do not happen among HC users. This may, simply put, be highly relevant for the many women who use HC and who might consider the absence of ovulatory cycle shifts desirable or undesirable side effects (Alvergne & Lummaa, 2010).

The present study

In the present study, we sought to replicate and extend central findings on cycle shifts in extra- and in-pair sexual desire, attractiveness, clothing choices, mate retention, and competitiveness, while also improving on methodological shortcomings in the cycle research literature. We preregistered our study and analysis plan before data collection to reduce our own researcher degrees of freedom and thereby the risk of false positives. We collected demographic, personality, and relationship data in an online intake survey. Participants were told that the study was about "relationship dynamics." Women were then invited to an online diary with up to 40 days per woman. After the diary, women completed a short follow-up survey about potential menstrual cycle disruptions and about their next menstrual onset after the diary. This design increased our power to detect any effects. It also allowed us to obtain prospective daily reports of menstrual onset, avoiding recall error, and to do backward-counting from next onsets, decreasing error in the estimation of conception probability. Because diaries were filled out on participants' personal electronic devices we could assess women's reported behaviour and experiences close to actual behaviour in both place and time. We automated the study process, decreasing our own ability to influence women's participation and responses. Because there was no cost per participant we recruited a large sample and included women regardless of contraception status, providing both a quasi-control group and making it less clear to participants what we were studying. We also assume that the automated, encrypted, minimal-contact online study made women feel more anonymous and hence comfortable to report, for instance, extra-pair desire and sex.

However, using this approach implied that we could not directly measure hormones, obtain photos of women, or collect ratings by their partners.

Because there is little agreement on best practices and standard operating procedures for doing this kind of research (Blake et al., 2016; Gangestad et al., 2016; Gildersleeve et al., 2014b), we also used a variety of robustness checks to test the consequences of different decisions during data processing and statistical modelling, especially conception probability estimation, exclusion criteria, and control variables.

Preregistered hypotheses

We registered the following hypotheses on the Open Science Framework (OSF) on the day that data collection began. We reworded and reorganised them slightly here for space and clarity (see OSF for original **osf.io/3ytjg**).

H1. Ovulatory changes (increases during fertile window among naturally cycling women in a heterosexual relationship, but not for hormonal contraceptive users) occur in
  - H1.1. female extra-pair desire and behaviour
  - H1.2. female in-pair sexual desire
  - H1.3. having and initiating in-pair sexual intercourse (if circumstances allowed, e.g. partner was close by)
  - H1.4. subjective feelings of attractiveness
  - H1.5. choice of clothing (self-rated on the dimensions “sexy,” “figure-hugging,” “seductive”)
  - H1.6. reported male partner mate retention tactics
  - H1.7. narcissism on both dimensions of the NARC (admiration and rivalry)

H2. Moderation or *shift* hypotheses: The ovulatory increase in women’s extra-pair desires and reported male mate retention behaviour is strongest (and the in-pair desire increase is weakest) for women who perceive their partners
  - H2.1. as low in sexual and physical attractiveness
  - H2.2. as low in sexual attractiveness relative to long-term partner attractiveness
  - H2.3. as less attractive compared to themselves

H3. Predicted ovulatory changes are larger than, and independent of, potential ovulatory shifts in self-esteem.

In addition, we preregistered to test trait differences in extraversion (H4.1.), shyness (H4.2.), and neuroticism (H4.3.) as potential ovulatory change moderators. We called these moderators exploratory in the preregistration to differentiate them from those already tested in the existing literature. As preregistering exploratory tests is not philosophically consistent, we would now call

these low-confidence predictions. We expected that the ovulatory increase in extra-pair desire (e.g. desire to attend social gatherings where they might meet men) may possibly be stronger for extraverted/outgoing than for introverted/shy women. Further, we expected that neuroticism may influence strength of the ovulatory increase in extra-pair desires and subjective feeling of attractiveness, though we did not specify a direction (H4.4.).

Methods

Power analysis

Because we used multilevel analyses for our within-subject data, we conducted simulations to assess our study's statistical power. We simulated data under a number of different scenarios, varying for example the effect size associated with conception probability, the sample size, the number of days sampled per participant, the standard deviation of the day of the ovulation (i.e. by how much our estimated conception probability missed the correct day on average), the trait component of the outcome, and whether participants were scheduled for sampling on predicted fertile vs. non-fertile days or on random days. We did not simulate between-subjects analyses, because these should be avoided not only because of their low power (Gangestad et al., 2016) but also for reasons of validity (see *Methodological issues*).

Researcher degrees of freedom simulation

Because researcher degrees of freedom have been discussed as a source of problems in the literature, we repeated our power analysis with an effect size of zero and assumptions simulating a hypothetical researcher engaging in the following *questionable research practices*: a) optional stopping (stop 20 or 10 participants earlier if $p < .05$), b) control for an irrelevant covariate if $p > .05$, c) try up to five correlated measures of a construct as outcomes, d) start with a continuous predictor, then try broad and narrow window if $p > .05$ and combinations of these practices and determined the number of false positives.

Preregistration and Ethics

Research that only entails self-reports is exempt from ethics committee approval under German regulations. We preregistered[1] our study's hypotheses and methods on March 19, 2014 and added a planned amendment[2] to our exclusion criteria and fertility estimation method to the preregistration on May 10, 2014, when data collection was already underway (Schilling, Straus, Arslan, Gerlach, & Penke, 2014). Participants enrolled from March 19, 2014 to July 2, 2015. The last diary entry was made on December 3, 2015. We did not preregister a fixed sample size, as this is hard to control in online studies and power analyses based on a biased literature are of limited use. Instead, we preregistered that we would obtain at least 200 eligible naturally cycling participants and wrap up data collection once we were unable to find further participants or after 18 months.

Participants

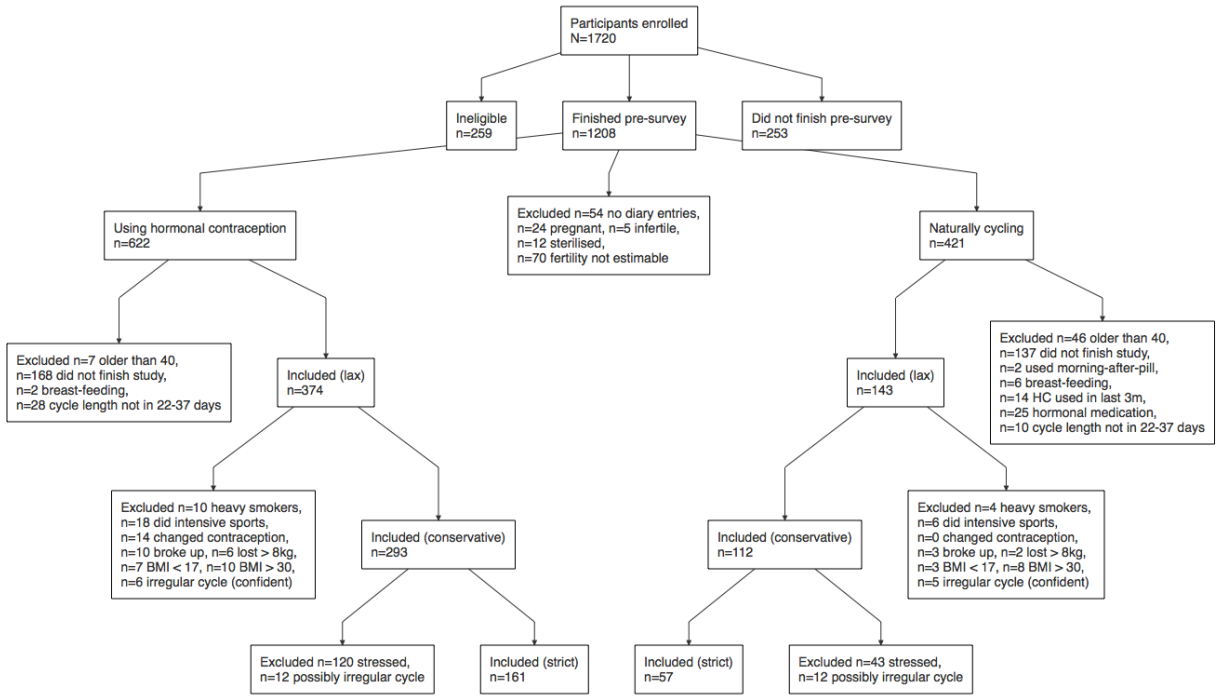
We recruited women via university mailing lists in Germany, Austria, and Switzerland, newspaper articles about our group's work (without references to ovulation-related work), our online study site psytests.de, word-of-mouth, and among local students in exchange for course credit at Georg August University Göttingen. Only participants who self-reported their sex as female and reported currently being in a heterosexual relationship were allowed to participate. Out of the 1,720 participants who signed up for the study, 259 were ineligible to participate according to these criteria, 253 did not complete the demographics and personality survey preceding the diary, 54 completed no diary entries, 41 were sterilised, infertile or pregnant, and fertility was never estimable for 70 due to menstruation never being observed or because of few or patchy diary entries. Out of the remaining participants, 60% (n = 631) were using some form of hormonal contraceptive. Of those, 88% used the pill, with most of the others using a vaginal ring or an intra-uterine device (IUD). A total of 40% (n = 428) were naturally cycling. Specifically, 5% (n = 53) used a fertility-awareness-based method, 28% (n = 291) used only barrier methods, mostly condoms, and 6% (n = 67) reported no contraception. 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 over-exclusive, as only 13% of the naturally cycling sample would have been retained. Hence, we differentiated our exclusion criteria into four strictness levels

[1] The preregistration had a second part, which pertained to hypotheses related to estrogen dosage effects in hormonal contraceptives and which we plan to discuss in a separate manuscript.

[2] In our initial preregistration, we specified that we would use backward counting from the observed next menstrual onset to estimate a narrow fertile window (reverse cycle days 15-19 vs. 2-11). After the publication of Wood et al. (2014), we amended the preregistration to *additionally* test a broad window (reverse cycle days 14-22) to compare results using the two windows. Moreover, we preregistered that we would descriptively show results based on continuous curves centred on the estimated day of ovulation.

and examined the effect of applying these levels in robustness checks. The participant flow and exclusion criteria are shown in Figure 1.

Figure 1. Participant flow. The figure depicts the various exclusion criteria and the number of participants affected by each (if not already excluded for a preceding reason).



The 1,043 eligible participants were on average 25.5 years old (*SD* = 6.3, range 18-53) and had been in a relationship for 3.8 years (*SD* = 4.3). Most (71%) were students, 24% were working, 3% were not working or described themselves as homemakers, and 3% were in secondary or vocational school. A majority reported their religious denomination as Christian (56%) and 42% described themselves as nonreligious. Twelve percent were married and a further 4% were engaged. Four percent of the sample reported not yet having had sex with their current partner. Most (88%) had no children. The largest group co-habited with their partner (41%), but a sizeable fraction had a long-distance relationship (31%), with the remainder living in the same city as their partner. Of those who did not live with their partner, 34% lived in a flatshare and 25% lived alone. We present more detailed data on the distance between partners, how often they saw each other and spent the night on the supportive website (**osf.io/pbef2**). Geographically, only our university town seemed visibly overrepresented. Hormonal contraceptive users differed from naturally cycling women in several ways (see Table 1 for continuous variables and supportive website for all others). Most importantly, they were almost 5 years younger on average, and consequently more likely to be unmarried and not to co-habit, to be in relationships for a shorter time (approximately 2 years), to have had 3.5 fewer lifetime sexual partners, to be students, and to have lower income. However, when simultaneously predicting hormonal contraception status from 28 demographic and personality predictors in a probit

regression, only lower age, lower openness, higher conscientiousness, and being unmarried were significantly predictive at $p < .05/28$. For the sample used in our preregistered analyses, the only differences remaining significant in the regression were that women on the pill were approximately 3 years younger and lower in openness. Hormonal contraceptive users also had shorter and more regular cycles, which might be consequences rather than causes.

Table 1. Descriptive statistics by hormonal contraceptive use.

	Mean (Standard deviation)			
Variable	**HC user**	**Cycling**	**Hedges' g**	***p***
Age	23.6 (4.4)	28.4 (7.6)	1.10	< .001
Religiosity	2.0 (1.1)	2.0 (1.2)	0.01	.891
Age at first time (years)	16.9 (2.3)	16.9 (2.4)	-0.01	.886
Age at menarche (years)	13.0 (1.3)	13.0 (1.5)	-0.06	.557
Relationship duration (years)	2.9 (3.0)	5.0 (5.5)	0.70	< .001
Cycle length (days)	27.9 (2.9)	29.1 (3.6)	0.41	< .001
Life no. sexual partners	5.7 (7.2)	9.3 (14.9)	0.50	< .001
BFI Extraversion	3.5 (0.8)	3.5 (0.8)	0.03	.638
BFI Agreeableness	3.6 (0.6)	3.6 (0.6)	0.00	.964
BFI Neuroticism	3.1 (0.7)	3.0 (0.8)	-0.14	.037
BFI Conscientiousness	3.6 (0.7)	3.5 (0.7)	-0.15	.024
BFI Openness	3.6 (0.6)	3.8 (0.6)	0.31	< .001
Relationship satisfaction	4.2 (0.7)	4.0 (0.8)	-0.20	.003

Notes. Constructs in bold remained significant after multivariate adjustment in a probit regression. BFI = Big Five Inventory. HC = hormonal contraceptive.

Procedure

Participants filled out web-based questionnaires on their personal electronic devices (27% used a mobile device). The study was implemented using the online open-source survey framework **formr.org** (Arslan & Tata, 2016). The software permitted us to automate all repetitive aspects of the study, such as administering surveys, sending email and text message invitations and to generate graphical feedback for participants. The study administrators communicated with participants through an email account and could send manual reminders and administer service requests in case of problems without seeing the participants' data.

Intake form and consent

First, participants were informed that the study's purpose was to examine the relationships between everyday life, relationship events, psychological well-being, and sexual behaviour. They were told that each diary day they filled out would add one more lot in a lottery for four Amazon.com coupons worth 20€ each and that they would receive extensive feedback on their personality and the longitudinal co-development of their mood, self-perceptions, and clothing choices over weekdays. Students of our university could earn course credit instead. They were informed that, although the study required their email address to send diary invitations, data would be stored separately and anonymously and that the feedback would also be generated anonymously and automatically.

Demographic and personality survey

After obtaining consent, we asked participants for their sex, age, and relationship status. Only self-identified females in a heterosexual relationship could proceed. Next, the women reported various demographics, details about their relationship, their menstrual cycle and contraception status and completed several measures of personality, relationship satisfaction and jealousy (see Table 2).

Diary

On the next day and until at least 30 entries were obtained over a period of at least 40 days, women were invited to fill out the diary via email and, if possible, text message at 5 pm Central European Time. They could fill out the diary until 7 hours after the invitation was sent. Participants completed the diary in a median time of 6.5 minutes. In each diary entry, they responded to 58 items about their relationship, interactions with their partner, clothing style, self-esteem, narcissism, sexual desire and behaviour, and menstrual cycle (see below). They were asked to refer to the time period since their last entry or 30 hours ago, whichever happened sooner. They could also give free-text responses to provide context for their entry.

Follow-up survey

After completing the diary (usually immediately after the last day), women were invited to a follow-up survey. In this survey, we asked several questions which we expected to relate to the validity of the results, namely what they thought the purpose of the study was, whether they were ill, took medication, lost weight, smoked, broke up with their partner, started a new relationship, switched contraception methods, or felt extraordinarily stressed. They then received their feedback. If they had not menstruated during the last 14 days of the diary, we sent them reminders every other day inviting them to tell us about their next menstrual onset, continuing until they did.

Measures

We documented all items for all surveys on the Open Science Framework (see osf.io/kd26j). To assess reliability for cross-sectional measures we computed Cronbach's alpha. For within-subject measures, we computed the generalizability of within-subject change aggregated across items (Shrout & Lane, 2012) using the *psych* package (Revelle, 2017). We document the main outcome measures for the diary and their reliabilities in Table 2. We used measures from previous studies where possible, but previous studies often could not or did not test the relevant generalizability metric for within-subject change. Unfortunately, the mate-retention-related measure did not appear to measure within-subject change reliably, and generalizabilities for the other outcomes were lower than optimal. The cross-sectional measures of personality, i.e. Neuroticism and Extraversion from the Big Five Inventory (Lang, Lüdtke, & Asendorpf, 2001) and shyness (Asendorpf & Wilpers, 1998), had Cronbach's *α*s ranging from .83 to .88. Confidence intervals (95%) for these *α*s had a width of 0.02-0.03. The reported physical attractiveness of the partner was based on two items (taken from Haselton & Gangestad, 2006) asking about his physical attractiveness and his sexiness (α = .80). The reported short-term attractiveness of the partner included the physical attractiveness scale, plus an item about his attractiveness for an affair or one-night stand and an item asking about sexual satisfaction with this partner (α = .62). To compute the partner's attractiveness relative to oneself (Haselton & Gangestad, 2006) we first computed a five-item mate value scale (Landolt, Lalumière, & Quinsey, 1995) for the partner and the participant. Own mate value (α = .73) correlated .17 with partner mate value (α = .69). We then tested whether the four-point Likert item "Who does better with the opposite sex? You or your partner?" favoured the partner most when his mate value exceeded hers. This was the case. Thus, we standardised and summed the mate value difference and the latter item (α = .76). The relative measure was almost uncorrelated with the various absolute measures (|r| < .07). Further details on scale construction and reliabilities are available on the supportive website (**osf.io/pbef2**). Confidence intervals (95%) for *α*s of the attractiveness-related scales had widths from .04-.07.

Table 2. Outcome measures in the diary.

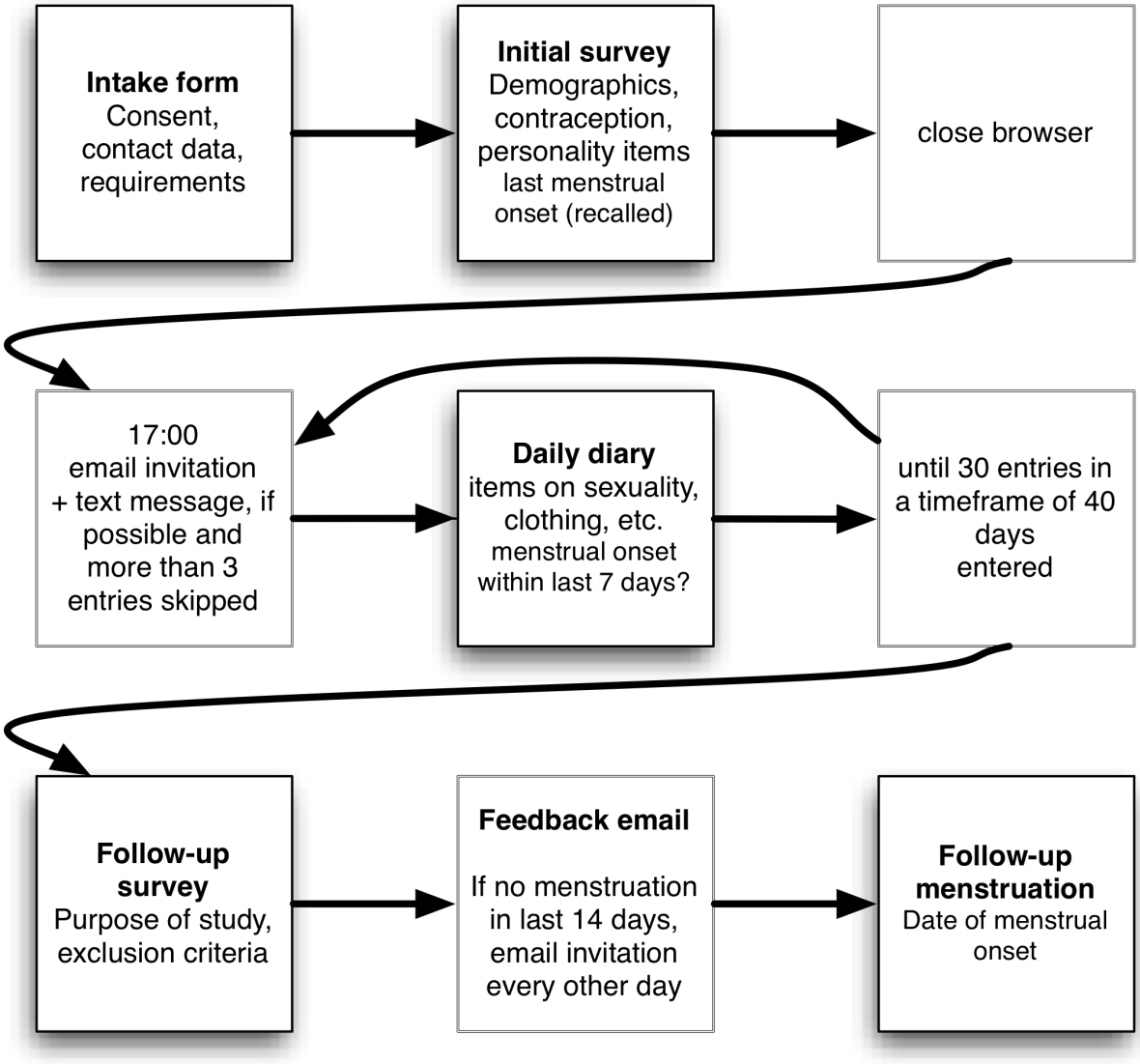
Construct	Scale Origin	Items	Rcn	Example item
Female Jealousy		3	.00	"I have asked my partner with whom he spent his day."
Relationship satisfaction		1	.85	"How satisfied were you with your relationship?"
"Sexy" clothing	Schwarz & Hassebrauck, 2008	3/8	.60	"Would you describe your chosen clothes today as sexy?"
Extra-pair desire	Haselton & Gangestad, 2006	12	.60	"I had sexual fantasies about men other than my partner."
Partner mate retention	Haselton & Gangestad, 2006	4	.00	„My partner asked me with whom I spent my day. "
Female mate retention	Haselton & Gangestad, 2006	6	.17	"I told my partner I love him."
Narcissistic admiration and rivalry	NARQ-K (Back et al., 2013)	3+3	.57/.55	"I felt worthy of being seen as a great personality."
Self-esteem	RSES Rosenberg, 1965	1	.86	"I was satisfied with myself overall."
Self-perceived desirability		1	.85	"I felt sexually desirable."
In-pair desire		3	.75	"I found my partner particularly sexually attractive."

Notes. Rcn = Reliability of change or generalizability of within-person variations. For clothing choices, three of eight items asked about "sexy" clothing choices.

Menstrual onset computation and fertile window inference

On each diary day, women reported whether they had had their period on that day or in the preceding 6 days. This meant women reported the same menstrual onset multiple times. Therefore, they could incorrectly recall a menstrual onset a few days later. To maximise accuracy, we always used the report closest to the reported onset. In all cases, women also reported a retrospectively recalled last menstrual onset in the survey preceding the diary. In some cases, women also reported a prospectively determined menstrual onset in a follow-up survey after the diary (see Figure 2).

Figure 2. Study structure. The figure shows the order in which the participants answered our questionnaires and where information about menstrual dates was gathered.



We used these dates to generate time series for each participant. We then counted forward and backward from each menstrual onset to the next or last menstrual onset, respectively. If the next menstrual onset was not available because women did not complete the follow-up survey, we could infer it from the reported average cycle length. We only used these inferred next onsets in our robustness checks. Irrespective of hormonal contraception status, we then computed a continuous estimate of the probability of being in the fertile window according to the method advocated by Gangestad et al. (2016), who based their estimates on Stirnemann, Samson, Bernard, and Thalabard (2013), among other data. This method accounts for the fact that the luteal phase length is less variable than the follicular phase and disattenuates the downward bias in the fertile window effect size that would result from uncertainty in estimating the day of ovulation and from potential anovulation.

Hormonal contraception users were assigned non-zero probabilities of being in the fertile window. We did this to rule out spurious effects unrelated to ovulation (e.g. distance to menstruation), using them as a quasi-control group in which our fertility predictor should have no effect. For our robustness checks, we used the continuous estimate of being in the fertile window; for our preregistered tests, we averaged the probabilities in a narrow and a broad window. Further details can be found on the supportive website (**osf.io/pbef2**). This procedure resulted in seven different fertility predictors, with varying number of diary days, see Table 3.

Table 3: The different conception probability estimates that were used as predictors.

Description	fertile window	n (days)	% of days	n (women)
all days		28,493	100	1043
narrow window, backward counted	15-19	9501	33.35	794
broad window, backward counted	14-22	11,497	40.35	796
narrow window, forward counted	11-15	12,171	42.72	973
broad window, forward counted	8-16	15,880	55.73	997
continuous, backward counted	n/a	17,614	61.82	817
continuous, backward counted from reported cycle length	n/a	26,580	93.29	1043

Notes. To make effect sizes across predictors comparable and disattenuate predictors for estimation error, we dummy-coded windowed predictors as being 0.053 on non-fertile days and 0.44 (broad)/0.51 (narrow) on fertile days. These were the averaged probabilities for those days from the continuous estimate, which varied from 0.01 to 0.58. Days were counted from the menstrual onset, starting at 1. The non-fertile window was defined as days 4-12 (backward-counted) or respectively days 18-26 (forward-counted).

Power analysis and researcher degree of freedom simulation

We documented our power analyses and researcher degrees of freedom simulations and results in more detail on the supportive website (**osf.io/pbef2**). They showed that under reasonable assumptions, power was a function of the number of usable days multiplied by the sample size.

To detect a regression coefficient of the fertile window of .2 with an alpha level of .01 in a sample of 150 naturally cycling women measured over 30 days, we had a power of .84 using a windowed predictor, because using windows meant not being able to use many of the measured days. Using a continuous predictor increased power to .99. In a sample of 500 women measured over 30 days, power approached 1. Power to detect an effect half/a quarter this size was still .97/.36 using a continuous predictor.

Data, code, results, and materials availability

We released all code, both for implementation and analysis, materials, and full statistical results pertaining to this study openly on the supportive website (**https://rubenarslan.github.io/ovulatory_shifts**; Arslan, 2018). We partially anonymised the data and uploaded them to the Open Science Framework for safekeeping (see **osf.io/kd26j**). However, because sexual diary data are hard to completely de-identify and extremely sensitive, we did not request consent from participants to share their data openly and cannot share the data publicly. Therefore, we can only share the partially anonymised data with anyone who has a valid reason and agrees not to attempt to re-identify the data. We have also generated a synthetic dataset using synthpop (Nowok, Raab, & Dibben, 2016). This dataset attempts to replicate many of the central features of our data, such as means and bivariate associations, but is anonymous. Others can use this to write code to test and build models using realistic fake data, which we can then easily replicate on the real data without having to vet them for access.

Analyses and Results

Preregistered tests

To test our hypotheses we fitted multilevel models in *lme4* (Bates, Mächler, Bolker, & Walker, 2014) with a random intercept per person and let our fertility estimate interact with a dummy for hormonal contraceptive use to predict the respective outcomes. Defining the model in this way allowed us to both test whether any ovulatory change among naturally cycling women was different from zero, as well as whether it was different from any changes occurring among hormonal contraception users. For Likert-scaled outcomes we fitted linear multilevel models and for categorical outcomes we fitted generalized linear multilevel models with a binomial family using a probit link. In Wilkinson notation (Bates et al., 2014, p. 4; Wilkinson & Rogers, 1973), the model equation can be formalised as

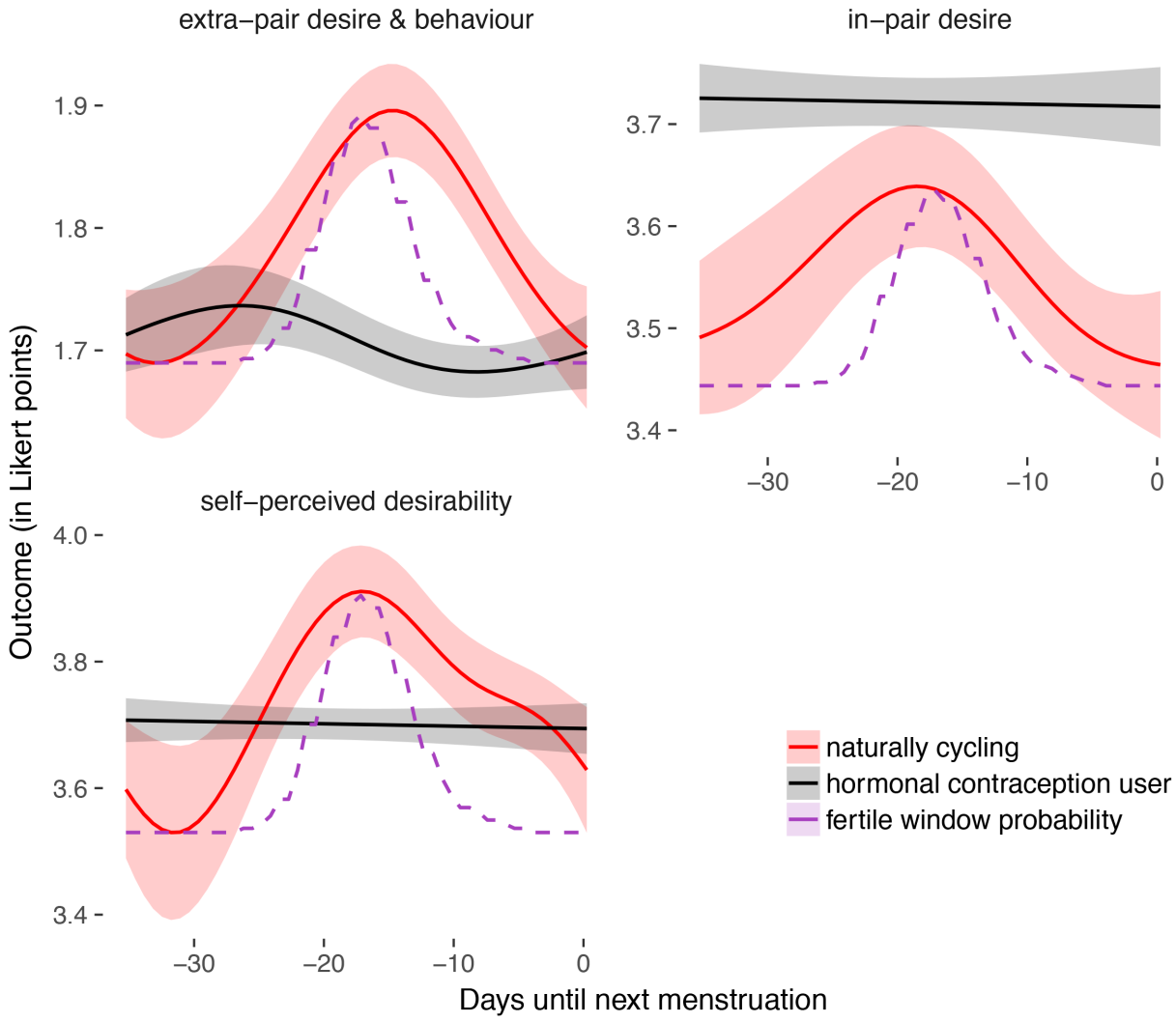
outcome ~ fertile_window * hormonal_contraceptive_user + (1 | person)

Here, *fertile_window* refers either to the backward-counted *narrow* or *broad* fertile window in the preregistered analyses. To test H3, we also refitted models with self-esteem as a covariate. Because we did not preregister it, we did not fit random slopes for the fertile window effect. We instead examine the effect of doing so in our robustness checks (Bates, Kliegl, Vasishth, & Baayen, 2015). To adjust for multiple comparisons, we set the significance level to .01 (see below). After applying our "lax" exclusion criteria (see robustness checks for further tests of stringency), we could use data from 143 naturally cycling women and 374 hormonal contraceptive users. Using the narrow (broad) fertile

window predictor, we could use 6,378 (7,740) diary days, or 12 (15) days per woman on average (see Table 3).

All outcomes are summarised in Table 4. For three outcomes, effects of the fertile window were significantly positive for naturally cycling women but absent for hormonal contraceptive (HC) users, a pattern we will refer to as *fertile window increases* in the following. When the interaction between HC use and the fertile window is of the same size as the fertile window effect, but negative, it indicates an absence of the change among HC users. We found small fertile-window increases in extra-pair desire and behaviour. Effects were significantly positive for all extra-pair subscales except the compliments subscale. We examined this pattern in more detail in the robustness analyses. Actual instances of intimate contact or sex with another person were very rarely reported (48 women reported extra-pair sex on 127 days, 112 women reported extra-pair intimate contact on 383 days), so that the log-odds-ratios seem large, but estimates were not significant ($ps > 0.17$). We also found small fertile window increases in in-pair desire, similar in size to the increase in extra-pair desire. On average, women did not have significantly more sex during the fertile window, but there were two consistent but only marginally significant moderators of the ovulatory increase in having sexual intercourse, namely cohabitation and average number of nights spent with the partner. Cohabitation moderated the changes, so that we observed no ovulatory increases among women in long-distance relationships ($p = .020$). Women who spent more nights per week with their partner also showed stronger ovulatory increases ($p = .048$). The increases were not stronger on the specific nights that the couple spent together ($p = .58$). Women did not initiate sex significantly more often in the fertile window. We also found small fertile window increases in self-perceived desirability, but not on wearing “sexy clothes.” The predicted effects were not significant for initiating sex, male mate retention, narcissistic admiration, and narcissistic rivalry (all $ps > 0.21$). As predicted, there were no significant effects on self-esteem and adjusting for self-esteem did not change other tested associations. The changes in self-perceived desirability, in- and extra-pair desire were also clearly apparent when plotting a smoothed spline over reverse-counted cycle days (Figure 2). The pattern of results held independently of whether we used a narrow or broad fertile window as the predictor.

Figure 3. Continuous display of outcome changes across the cycle. Smooth thin-plate splines (S. N. Wood, 2003) fitted over days until next menstruation with three central outcomes. The dashed line shows the estimated probability of being in the fertile window for each day. The shaded areas reflect 95% confidence bounds pooling days over participants for simplicity. To account for the cyclical nature of the data, we spliced in duplicates of the time series at both ends before estimating the splines and then dropped them afterwards.



None of the three main predicted moderators, i.e. the partner's short-term, sexual, and relative attractiveness, significantly exhibited the predicted pattern for any outcome (*p*s > 0.07), and some patterns went descriptively in the opposite direction of the prediction. Also, none of the personality variables moderated changes in extra-pair desire and behaviour (p*s* > .32). A test of whether neuroticism moderated shifts in self-perceived desirability was significant (p = .002), but inspection of marginal effect plots showed this to be driven by significant increases in desirability among highly neurotic hormonal contraceptive users, an unpredicted and likely spurious result.

Table 4. Preregistered associations, using the narrow fertile window

Outcome	Intercept	fertile	HC user	HC user x fertile
Extra-pair desire and behaviour				
extra-pair (EP) desire & behaviour	1.75±0.05	**0.27±0.06**	-0.05±0.06	**-0.30±0.07**
		$p < .001$	$p = .373$	$p < .001$
- EP compliments	2.37±0.08	0.25±0.11	-0.11±0.10	**-0.37±0.13**
		$p = .023$	$p = .267$	$p = .005$
- EP flirting	1.36±0.04	**0.15±0.06**	-0.09±0.05	**-0.22±0.07**
		$p = .006$	$p = .078$	$p < .001$
- EP going out	1.99±0.09	0.24±0.15	0.24±0.10	-0.31±0.18
		$p = .113$	$p = .019$	$p = .088$
- EP sexual fantasies	1.50±0.06	**0.49±0.09**	-0.19±0.08	**-0.43±0.11**
		$p < .001$	$p = .012$	$p < .001$
- EP desire	1.65±0.05	**0.34±0.06**	-0.13±0.06	**-0.31±0.07**
		$p < .001$	$p = .047$	$p < .001$
extra-pair intimacy[pb]	-4.47±0.30	0.89±0.42	-0.22±0.37	-0.57±0.72
		$p = .033$	$p = .554$	$p = .431$
extra-pair sex[pb]	-4.60±0.39	0.60±0.56	-0.44±0.57	0.17±1.08
		$p = .282$	$p = .444$	$p = .873$
In-pair desire and behaviour				
in-pair desire	3.48±0.08	**0.31±0.12**	0.24±0.09	**-0.39±0.14**
		$p = .010$	$p = .010$	$p = .008$
sexual intercourse[pb]	-0.98±0.07	0.12±0.17	0.17±0.08	-0.26±0.20
		$p = .483$	$p = .026$	$p = .203$
sex initiated by partner[pb]	0.26±0.09	-0.14±0.31	0.12±0.11	0.11±0.37
		$p = .642$	$p = .276$	$p = .775$
partner mate retention	2.86±0.07	0.05±0.09	0.00±0.08	-0.12±0.11
		$p = .569$	$p = .954$	$p = .255$
Self-perceived desirability and clothing choices				
self-perceived desirability	3.72±0.08	**0.37±0.13**	-0.07±0.09	-0.38±0.15
		$p = .004$	$p = .477$	$p = .012$
sexy clothing	3.16±0.07	-0.14±0.10	0.02±0.08	0.09±0.12
		$p = .169$	$p = .831$	$p = .492$
Narcissism				
narcissistic admiration	2.69±0.10	-0.05±0.08	-0.14±0.11	-0.09±0.09
		$p = .551$	$p = .214$	$p = .335$
narcissistic rivalry	1.29±0.04	-0.03±0.05	0.05±0.05	-0.02±0.06
		$p = .535$	$p = .322$	$p = .747$

Notes. Coefficients significant at $p < .01$ (before rounding) are bold. Associations with outcomes marked [pb] were estimated in a probit regression. The number after the ± is a standard error. Scales starting with EP are

subscales. The sex initiation item asked whether it was rather the partner or rather the participant who initiated sex, in a forced-choice question. Positive effects reflect that it was rather the partner.

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 supportive website, **osf.io/pbef2**).

Robustness checks

To test our results for robustness, we used a variety of approaches. Given the wide-ranging exploration and varying questions asked across outcomes and models, a null hypothesis testing approach would have been inappropriate. Instead, we focused on visualisations and the fertility effect's point estimate and confidence interval. We inspected effects to look for evidence that an effect was not robust (i.e. shifts in estimates that might not be explainable by sampling error). Here, we verbally and visually summarise the most important patterns. The checks are described more extensively in Supplementary Table 1. Our results are fully documented on the supportive website (**osf.io/pbef2**).

First, we built a baseline model that deviated from our preregistered procedure but implemented the best practices that were published after we preregistered (Blake et al., 2016; Gangestad et al., 2016). Here, the probability of being in the fertile window was continuously estimated from backward counting from the next menstrual onset, according to Gangestad et al. (2016). In cases where the next menstrual onset was not observed, we fell back to the next menstrual onset inferred from the average cycle length that women reported in the screening survey (see Table 3). Because using a continuous predictor means that days on which women were menstruating or in the premenstrual phase were also included, we included dummy variables for the reported menstruation and the inferred premenstrual phase (the six days before the menstrual onset). We also adjusted for the average probability of being in the fertile window per woman as an additional predictor, to ensure within-person estimates (Bafumi & Gelman, 2006). We let our fertility and menstruation predictors interact with hormonal contraception status, as in the preregistered tests. In Wilkinson notation, the model can be formalised as

outcome ~ (*fertile_window_probability* + premenstrual_phase + menstruation) *
hormonal_contraceptive_user + average_fertile_window_probability+ (1 | person)

In this baseline model, we included all usable data (from 1,043 women, 421 naturally cycling) instead of excluding many women based on our preregistered criteria. This way, we were able to include 25,948 diary days, i.e. on average 25 days per woman and more than 3 times as many days as in the preregistered analyses. We repeated all preregistered tests using this bigger dataset and the adjusted model. We then tested robustness by fitting numerous variations on the baseline model described above and examining the effect size and standard error of the fertile window predictor across models. We summarise what we consider the main patterns. Unless otherwise mentioned, results were robust to including more data and to the various checks described below. We conducted a total of 39 robustness checks, in five broad groups, per outcome. We abbreviated them by group and number (e.g., *M_c2* for second covariate model). With these abbreviations further details can be found on the supportive website (**osf.io/pbef2**) and in Supplementary Table 1.

In model *M_r1*, we allowed a varying slope per participant for the fertile window and the two menstruation dummy variables, a "maximal" specification that is somewhat controversial because of the potential for overparameterisation (Barr, Levy, Scheepers, & Tily, 2013; Bates et al., 2015). The random slopes for the fertile window predictor were substantial: larger than for the menstruation predictors and as large as the residual variation and the variation explained by the random intercept. However, accounting for random slopes did not change any conclusions about the average effect of the fertile window.

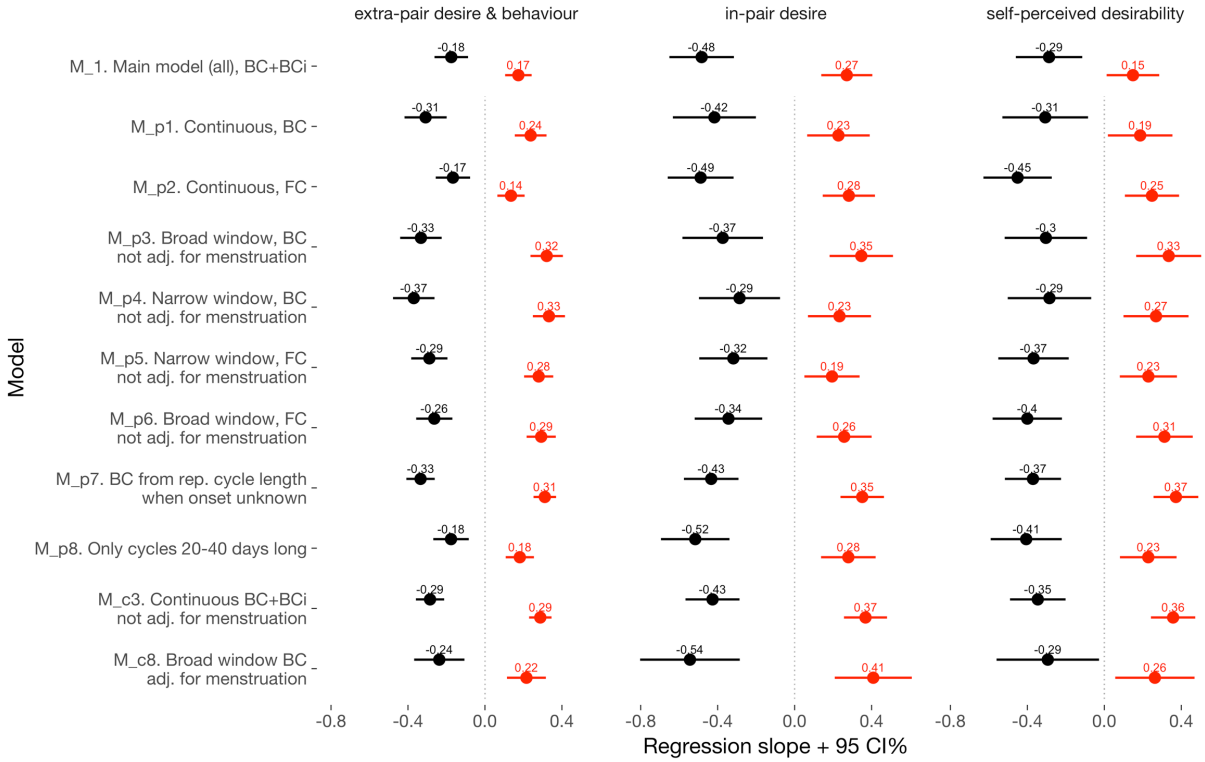
Exclusion criteria

We tested four levels of stringency for exclusion ("all," "lax," "conservative," "strict," see Figure 1) in models *M_e1-4* and *M_m5*. The stringency of our preregistered exclusion criteria, designed to exclude women with potentially anovulatory cycles, did not moderate the effect sizes in the expected way, i.e. we did not observe that effects became stronger with more stringent criteria. When applying stricter exclusion criteria, some confidence intervals overlapped with zero, but this seemed to reflect the heavily decreased sample size (see Figure 1). We also tried to implement a post-hoc criterion (*M_e5*) for data reliability, under which we excluded 1251 diary days (4% of all) where participants a) gave the same answer to all Likert items (n=23), b) accessed the diary later or earlier than intended due to technical problems (n=896), or c) took more than 24 hours (n=376) or less than a minute (n=30) to finish filling out the diary. We took these steps to reduce the number of careless responses and to remove days on which the assigned cycle day might be off. However, excluding these days had no noteworthy effect. We also tested (*M_e6*) the effect of excluding women who were trying to get pregnant, an exclusion criterion we had not preregistered. This exclusion attenuated the effect on in-pair desire, but did not eliminate it, and did not change results otherwise.

Different predictors and the within-subject method

In models *M_p1* to *M_p11*, we tested different estimates of the fertile window as our predictor to address the concerns about varying standards described in *Methodological issues*. We compared all combinations of a narrow window, broad window, continuous estimates, and backward- and forward-counting. When we used a continuous fertile window predictor, we also adjusted for premenstrual and menstrual days. We found that including adjustments for menstruation and pre-menstruation (*M_c3*) reduced effect sizes for the fertile window predictor. We could not always adjust for (pre-)menstruation when using a narrow window predictor because of model convergence problems. After taking this into account, we found no systematic pattern in which certain predictors (narrow or broad window, forward or backward counted) had larger effect sizes than others across outcomes (see Figure 4). However, continuous curves over backward-counted days (Figure 3) matched the predicted pattern more closely than curves over forward-counted days (see supportive website, **osf.io/pbef2**).

Figure 4. Robustness checks for predictors. Coefficient plot showing a consistent effect of the fertility predictor among naturally cycling women (red) but not hormonal contraception users (black) across several predictor and model specifications (explained in further detail in the text). FC = forward-counted from last menstrual onset, BC = backward-counted from observed next menstrual onset, BCi = backward-counted from inferred next menstrual onset.



To address concerns about between-subject studies and statistical power (see *Methodological issues*) empirically, we then tested whether effects could be shown using only a single day per participant (*M_d1*), two days (at low and high fertility; *M_d2*), four days (two each; *M_d3*), or by averaging high and low fertility days (*M_d5*). We found that none of the associations that were significant in the pre-registered analyses would have been discovered had we used between-subject analyses or a high-low fertility within-subject design with only two days.

Importance of covariates

To transparently show how much modelling decisions that might be considered researcher degrees of freedom matter, we fitted models *M_c1* to *M_c5*. In these, we added adjustments, one model at a time, for *M_c1* self-esteem, *M_c4* weekday and week number, and *M_c5* the time when the diary was started and how long it took to fill out, or we omitted adjustments for *M_c2* average fertile window probability, or *M_c3* both average fertile window probability and menstruation. This allowed us to see the effect these adjustments had on the estimated fertility effect. In *M_c6* to *M_c7*, we tested two different temporal autocorrelation models as opposed to the unstructured random effect correlations in our main model. In M_c9, we tested whether measurement reactivity might confound our results, by adjusting for splines for the number of days since the diary began (one variable for days filled out and one including missing days), by hormonal contraceptive use. Except for omitting the adjustment for (pre-)menstruation, none of these decisions led to substantial changes.

Different contraceptive methods

To partially address issues of reproductive intentions (see *Methodological issues*), in *M_m1* we compared four groups of contraceptive methods (hormonal, awareness-based, barrier-based, none). For women who combined multiple methods, the order of the list above determined precedence. The pattern of results was complex. The ovulatory increase in extra-pair desire tended to be larger for fertility-aware women (5% of the sample) and this was not merely because they had more regular cycles. Still, women using barrier methods or no contraception also showed ovulatory shifts. By contrast, the changes in in-pair desire and self-perceived desirability appeared weaker or absent in fertility-aware women, but stronger in women using no contraception (6% of the sample). Because women using methods other than hormonal contraceptives and barrier methods made up only a small minority of the sample, we could not rule out sampling variation as an explanation.

Methodologically relevant moderators

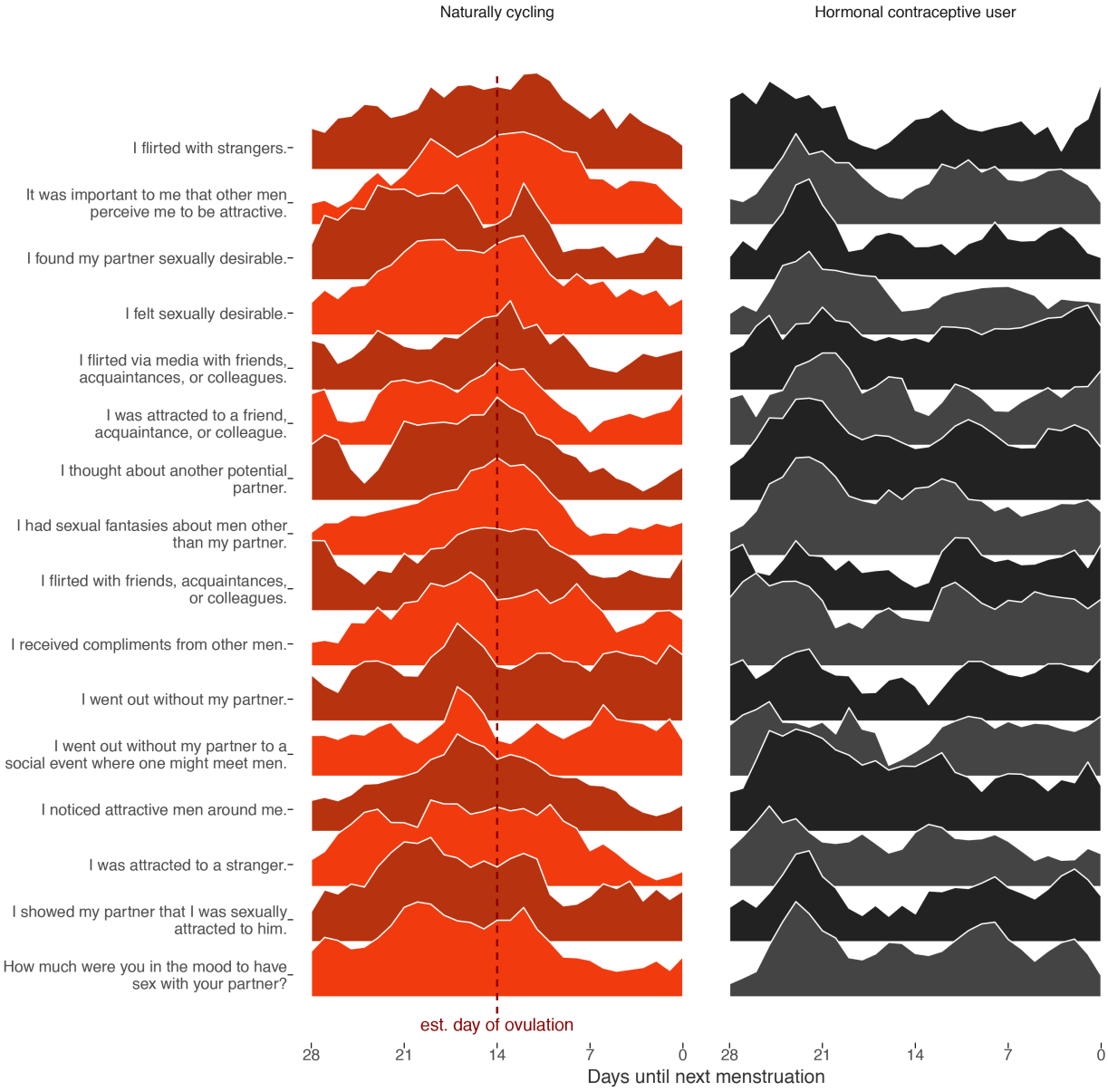
We then tested various moderators to prod different methodological issues. To test generalizability, we tested moderation by participant age (in groups 18-20, 20-25, 25-30, 30-35, 35-45, 45 and older, *M_m2*), and whether the weekday (*M_m3*) or the weekend (*M_m4*) moderated effects (Roney & Simmons, 2013). Except for age these moderators had no effects. Older individuals seemed to show stronger in-pair desire changes. Because the validity of fertility estimates from counting methods depends on accurate reporting and regular cycles, we tested for moderation by cycle length (*M_m6*), by self-reported certainty about menstruation parameters (*M_m7*), and by self-reported cycle regularity (*M_m8*). To further test generalizability, we also tested for moderation by cohabitation (*M_m9*) and by marital status (*M_m10*). Across outcomes, effects tended to be largest for women with cycle lengths between 25 and 30 days, and for women who were more certain about their menstruation parameters, but not for women whose cycles were more regular in general.

Differences across items and scales

We also ran Bayesian regression models using Stan (Bürkner, in press; Carpenter et al., 2015) to be able to appropriately model the positively skewed, ordinal distribution of the Likert items for extra-pair desire (i.e. many respondents indicated minimal extra-pair desire) in an ordinal regression using a cumulative outcome distribution and random effects for items and participants. In the Stan models, we also tested for heterogeneity of effect sizes across participants and items. In additional Stan models, we fitted a thin-plate regression spline (S. N. Wood, 2003) over backward-counted cycle days to examine whether the continuous probability of being in the fertile window would be a good fit to the shape of the estimated effect. In exploratory analyses, we also fitted one Stan model per item and graphically summarised the posterior densities for the conception probability estimates. Because of computational limitations, we fitted models separately instead of pooling information across items and scales.

Inspecting time series of within-subject change by item (Figure 5) for the three outcomes that were significant in the preregistered analysis, namely extra- and in-pair sexual desire and self-perceived desirability, showed that naturally cycling women tended to exhibit peaks around the estimated day of ovulation, while hormonal contraceptive users exhibited no clear peaks or minor peaks around menstruation.

Figure 5. Item-by-item plot of within-subject change. The trails in this plot represent within-subject change as a percentage of the maximal peak. Plots are smoothed with a moving average over three days. Items are ordered top to bottom by how late in the cycle the highest peak occurs for naturally cycling women.



We also attempted to test whether the effects of in- and extra-pair desire were different in size and independent of each other, to test whether they were potentially both driven by a third variable, such as increased target-unspecific sex drive (see supportive website, **osf.io/pbef2**).

Robustness checks summary

With the additionally included data, estimates of fertile window increases in extra-pair desire and behaviour, in-pair desire, and self-perceived desirability were robust, but standard errors shrunk to about half. Apart from this, the overall pattern favoured backward counting and including as much data as possible using continuous predictors, but windowed and forward-counted predictors performed similarly. More importantly, when we adjusted for (pre-)menstruation, estimated fertility effects were often reduced. Further, none of the predicted moderation patterns turned significant when adding more women, and using slightly different items for the partner attractiveness moderator variables did not change the pattern. No fertile window increases emerged for any other outcomes, including further outcomes for which we had not predicted increases (e.g., female jealousy, relationship satisfaction). We found increased effect sizes for some combinations of outcomes and contraception methods, but had only small sample sizes for any methods except hormonal contraception and barrier methods. Effects were also stronger for women with average length cycles and for women who were more confident in their own cycle parameters estimates. We also found that older women and women who were trying to conceive experienced stronger increases in in-pair desire in the fertile phase, but again had only limited sample sizes for these groups.

Discussion

In the present large diary study, we aimed to replicate reports of ovulatory changes in extra- and in-pair sexual desire and behaviour, as well as related outcomes, and to test several methodological concerns. We could replicate only some of the previously reported ovulatory changes, namely those in extra-pair sexual desire and behaviour, in-pair sexual desire and behaviour, and self-perceived sexual desirability. In Figure 3, we show that changes across the cycle for these three outcomes closely match the probability of being in the fertile window.

Main effects of the fertile window

Extra-pair desire and behaviour

We found robust support for a fertile window increase in extra-pair desire and behaviour. This scale was a fairly heterogeneous aggregate of items measuring increased attraction to, fantasizing about, flirting with, receiving compliments from, and going out to meet with men other than the primary partner. In separate analyses, we also examined whether women were more likely to be intimate or have sex with other men during the fertile window. While descriptively supporting the predicted ovulatory shifts, these events were rare and effects were not significant. We also examined effects on

the subscale level. Fertile window increases in sexual fantasies were descriptively strongest, but the aggregation of subscales seemed justified.

In-pair desire and behaviour

We found robust support for fertile window increases in in-pair desire. Although in-pair desire predicted intercourse with the partner, ovulatory increases in sexual intercourse were not significant in our preregistered analyses. We may have had too little power to detect mean shifts in this dichotomous behaviour: Women reported sex on 21% of days and 67 women who filled out the diary on more than 25 days never reported sex with their partner at all. In our robustness checks that included more data we observed increases, but only in comparison to the HC group (which decreased non-significantly). Further, as predicted, two indicators of partner availability moderated the sexual intercourse shifts in the preregistered analyses marginally significantly: ovulatory increases were absent among women in long-distance relationships and among those who reported rarely spending the night with their partner. The daily report of whether the couple spent the night together did not moderate the shift, but the same-day behaviour may act as a mediator, not moderator, of ovulatory shifts in sexual behaviour. We see this pattern as partial support for our hypothesis 1.7., stating that ovulatory increases would be observed if circumstances allowed it (e.g., the partner was nearby). This pattern is consistent with the findings for coupled women in a larger within-subject study on 1,180 women and 37,170 diary days (Caruso et al., 2014), but runs counter to previous results from 20,000 women in a between-subject study (Brewis & Meyer, 2005). Shifts in in-pair desire also appeared to be stronger for women cohabiting with their partner, a pattern we did not predict.

Mate retention, jealousy

We observed no fertile window changes in partner mate retention, but the generalizability of change for these items was very low, making the detection of an effect unlikely. We based our items on the previous literature, in which generalizabilities of change were not reported. We had preregistered a suboptimal procedure for improving outcome reliabilities, based on assessing Cronbach's alphas, which ignore the multilevel structure of the data. Instead, we 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* behaviours, as opposed to *persuasive* behaviours, because items measuring the former seemed to show stronger changes.

Self-perceived desirability and clothing choices

We found fertile window increases in self-perceived desirability in our preregistered analyses that were robust to our checks, although standard errors were relatively broad because we used only a single item to assess this outcome. Contrary to our predictions, we found no fertile window changes in self-reported "sexy clothing," even though this was associated with desirability. As predicted, we also found no change in "flashy/showy" clothes and self-esteem in our robustness checks. These results are consistent with recent large-sample replications of fertile phase increases in facial attractiveness (Jones, Hahn, Fisher, Wang, Kandrik, Han, Lee, et al., 2017). Given their results the ovulatory changes in self-perceived attractiveness in our data might track direct hormonal effects on physical attractiveness (e.g., clearer skin) and not just reflect a change in self-perception.

Other outcomes

For all the other outcomes we found no ovulatory changes that were at the same time absent among HC users. Reassuringly, in no case did we observe any significant associations for outcomes for which we predicted none (relationship satisfaction, self-esteem, spending the night/communication with the partner, female jealousy, and female mate retention). Nor did we find associations for the narcissism outcomes, for which we had indirectly extrapolated our predictions from prior reports in the literature of ovulatory changes in clothing, interpreted as signs of intrasexual competition (Durante et al., 2008). We should reiterate in this context that we did not replicate cycle shifts on clothing choices either. Perhaps this can be interpreted as evidence that the literature suffers more from potential false positives than from false negatives, though it is noteworthy that some previous studies had not found ovulatory increases in in-pair sexual desire and behaviour, nor main effects of the fertile window on extra-pair desire (e.g., Brewis & Meyer, 2005; Haselton & Gangestad, 2006). We emphasize that in the current study both negative and positive results were largely robust to the many different analytic approaches that we tested.

Predicted moderator effects and individual differences

There was insufficient evidence for moderation of male mate retention behaviour, extra-pair, or in-pair desire by the partner's attractiveness (no matter whether it was assessed as relative to self, sexual, or physical), as predicted by the *good genes ovulatory shift hypothesis*. Although some patterns descriptively pointed in the predicted direction, none of the predicted patterns were significant, and some were opposite to our predictions. Because only 144 naturally cycling women remained for our preregistered analyses, statistical power may have been insufficient to detect plausible moderation effect sizes. However, we found no evidence for moderation effects in the more inclusive sample of our robustness tests either. Although our sample sizes are bigger than many

published studies that reported a moderation effect (Haselton & Gangestad, 2006; Pillsworth & Haselton, 2006b), we would ideally prefer to exceed their power by a wider margin due to the winner's curse, i.e., effect sizes being overestimated through selection and publication bias. We should also mention that some of the earlier studies we aimed to replicate (Haselton & Gangestad, 2006; Larson et al., 2013) did not actually report significant main effects of the fertile phase. Increases were reported to be *qualified* by a moderator. In this sense, we replicated neither findings on main nor on moderator effects from these studies. Still, we believe GGOSH can be taken to predict main effects as well, because amplified shifts in some women whose partners lack certain characteristics should, averaged across women, still yield detectable main effects. There are some conceptual similarities between ovulatory shift moderators of extra- and in-pair desire and direct tests of ovulatory changes in mate preferences, because both are shifts in who is preferred as a mate. Newer, more adequately-powered laboratory research also sheds doubt on ovulatory shifts in preferences for facial masculinity (Jones, Hahn, Fisher, Wang, Kandrik, Han, Fasolt, et al., 2017). All in all, our findings do not support GGOSH, as we find no substantial moderator effects by partner attractiveness.

We found no evidence for the tentatively predicted moderation of increases in extra-pair desire or self-perceived desirability by neuroticism, extraversion, or shyness. However, because we had on average 25 days for each woman, we could estimate inter-individual differences in ovulatory increases (i.e. random effects for the fertile window). Random effect variances for the fertile window predictor were substantial. Hence, there might be real heterogeneity in ovulatory increases to be explained. Future research should test and improve the reliability of these inter-individual differences across cycles.

Theoretical implications

Although further tests should be conducted, the *good genes ovulatory shift hypothesis* could be wrong, given that we could not replicate previously reported moderators. More recent theoretical work emphasises that predictions of adaptive extra-pair sex, which Pillsworth and Haselton (2006a) call dual mating, should be divorced from predictions of ovulatory changes in mate preferences that do not necessarily precipitate extra-pair sex, but still function to bias sire choice (Gangestad, Thornhill, & Garver-Apgar, 2015). We cannot test all aspects of these recent theoretical developments in our study. An alternative, simpler explanation (Roney & Simmons, 2013) is based on life history theory. It suggests the observed increase in sexual desire during the fertile phase reflects a motivational priority shift towards reproduction. The purported function would be to accept higher costs of sex (such as energetic and opportunity costs, or sexually transmitted infections) the more likely it is that sex leads to conception. This theory also predicts fertile phase drops in somatic investment, such as food intake (Fleischman & Fessler, 2007; Roney & Simmons, 2017). In this study, we did not assess any non-reproductive motivations. Testing whether the effect on extra-pair desire is bigger than the one on in-

pair desire, which could potentially distinguish preferential shifts (sire choice) from motivational shifts across the ovulatory cycle, is also not possible in our data, because we did not use parallel items to assess both and did not assess single women. Effect sizes were descriptively similar (see supportive website for an extended discussion of these issues, **osf.io/pbef2**). Item-level comparisons showed that effects were on average larger for items that required no object of desire to be present and no action to be taken (e.g., "I had sexual fantasies about men other than my partner"). Jones et al. (2018) faced the mirror image problem: in their data, they found effects of estradiol and progesterone on generalized and solitary sexual desire, which were not moderated by relationship status, but they could not differentiate between in- and extra-pair desire. Ultimately, we currently cannot tell whether general, target-unspecific sexual motivation drives the effects on in- and extra-pair desire we find (Roney, 2009). Roney and Simmons (2016) report that there seem to be stronger extra-pair desire changes among partnered than among single women, but their sample of partnered women was very small (n=8). Future studies should be designed and powered to discriminate between theories. Relatedly, theoreticians should make exact predictions down to what certain statistical models will find, because verbal ambiguity might otherwise preclude the identification of the best supported theory.

Hormonal contraception

Whenever we found an ovulatory increase, we also found that it was absent among hormonal contraceptive users. In this sense, we identified one reliable moderator. The absence of these cycle changes probably reflects the suppression of ovulation and concurrent hormonal changes. Moreover, estimated effects of menstruation and the premenstrual phase on psychological outcomes, as measured in the diary, were also attenuated among HC users. In the preregistered analyses, we found only small and statistically non-significant mean level differences between HC users and cycling women in the diary outcomes, as well as in the demographic and personality variables that we tested. These differences are presumably confounded by selection and attrition effects. For example, women who expect their relationship to last may be more likely to start using HC and to show less extra-pair desire, and women who experience libido decreases on HC may go off it again. Thus, the (absence of) mean level differences may not (entirely or at all) speak to causal effects of HC.

There are few randomised controlled trials (RCTs) that can answer questions about psychological changes caused by HC use. Existing ones so far mostly ignore cycle phase (Zethraeus et al., 2016, 2017, but see Ranehill et al., 2017) thus not yielding the full picture of differences across the cycle. Potentially, this can lead to spurious or misleading conclusions of differences, if women in the naturally cycling control group are measured in different cycle phases across time points. As the effects of cycle phase on sexual desire in our study were similar in size to effects reported for

hormonal contraceptives in Zethraeus et al. (2016), future RCTs should always tease cycle phase and HC influences apart.

The suppression of cyclical psychological changes is not currently being pointed out as a side effect of the pill in package leaflets, although they do mention potential effects on libido and appetite. Potentially, decreased fluctuations in extra- and in-pair desire might be seen as less worrisome than e.g. decreased average levels of libido or altered mate preferences (Alvergne & Lummaa, 2010), but ideally HC users should get the chance to make this decision themselves based on more complete information. Decision making about HC use may vary, e.g. some women may prefer to have cyclical ups and downs, while some may prefer to have a lower but constant mean level. Moreover, individual differences in the actual physiological and psychological response to HC may be more important than differences in preferences for side effect avoidance and should be a future research priority.

Limitations

In this study, we relied on self-report, which may mean that social desirability, measurement reactivity, and recall error could affect our results. We hope we succeeded in minimising these issues by ensuring privacy and anonymity for participants, preventing access to past responses, asking specific closed-form questions daily, and statistically testing and adjusting for temporal trends (Barta, Tennen, & Litt, 2012). Some women in this sample may have used fertility tracking apps as a supplemental contraceptive method or simply out of interest. Such women may not have reported using these apps, because we only asked about contraception. Potentially, the increased awareness of these women could have influenced our results. An obvious improvement would be to also collect partner- and potentially peer-reports, although this might have negative consequences for the perceived anonymity of responses. To decrease measurement reactivity and to test its effect, future studies could space out diary invitations over a longer period, for instance by sending them only on odd days or tailoring them to predicted (non-)fertile phases. Ideally, the schedule would be varied randomly by group (Barta et al., 2012).

We overestimated how conscientiously participants would fill out the diary. Hence, some women strung out the participation period over such a long time that menstruation could have occurred in an unobserved period, because women only reported menstrual onsets that occurred fewer than 7 days ago. Therefore, fertility was not estimable for ~6% of days (Table 3). Further, sending daily invitations via email presented a technical challenge. Due to delays in the sending process and spam filters some emails occasionally arrived a few hours late or not at all. We introduced text message reminders approximately halfway through the study and remedied this somewhat. These problems are presumably unrelated to outcomes and cycle position as *M_e5* shows, but still worth avoiding in the future. Because we required 30 complete daily reports before the study could end, some women

never concluded our study, leading to 31% dropout in the follow-up survey. Future studies should use a fixed timespan for the diary, so that the follow-up takes place at the same time regardless of participation rate.

We only asked participants whether they had been intimate with someone other than their partner, but failed to systematically ask about the context and sex of the person. Free-text responses showed that several instances of reported extra-pair activity were not cheating with another man, but polyamorous or open relationships, affairs with women, or sex with the partner and another couple or a third person. All of these have dubious relevance to the research question about adaptive benefits of extra-pair infidelity. We also did not collect data on single women, preventing us from discriminating between an increased propensity for flings in general versus extra-pair infidelity. Future work should also differentiate sexual activity more than we did here, including not just sexual intercourse and other sexual activity with the partner, but also masturbation and nonsexual intimacy.

The generalizability of change for our outcome scales was sometimes zero and in other cases suboptimal. Previous research, from which we derived our scales, may have suffered the same problem, but did not conduct the appropriate psychometric analyses to find out. We think menstrual cycle research should learn from work on psychometrics and measurement in personality development research (Shrout & Lane, 2012). Mirroring the old person-situation debate (Kenrick & Funder, 1988), the evolutionary literature now debates the relative importance of between and within person variation(Havlíček, Cobey, Barrett, Klapilová, & Roberts, 2015; Jones, Hahn, Fisher, Wang, Kandrik, Han, Lee, et al., 2017; Jones et al., 2018; Zietsch et al., 2015). However, without using improved methodology to study within-person variations the debate cannot be resolved (Roberts & Caspi, 2001; Shrout & Lane, 2012).

Our sample was a convenience sample. Although it included a broad range of women, many (73%) were students, most (87%) had no children, few (12%) were married and all spoke German. Generalizability to older and higher-fertility populations, especially from settings that are not western, educated, industrialised, rich and democratic (WEIRD; Henrich et al., 2010) is thus limited. Although we assume universal hormonal mechanisms drive our effects, average hormonal levels might differ substantially for women who do not cycle regularly, for instance because they have recently been pregnant or breastfeeding, or because they have worse nutritional status. Hormonal assays would help to better understand such patterns. In Western societies, female infidelity is not uncommon, with a 12-month prevalence of 2-4% and an occurrence of 20-25% per marriage (Fincham & May, 2017). However, few women have children with an extra-pair mate (1-2%; Larmuseau, Matthijs, & Wenseleers, 2016). Initial studies on ovulatory shifts were based on estimates of a higher extra-pair offspring rate, but even few instances may suffice to exert the necessary selective pressure. It has been suggested that the low rate is a evolutionarily recent, cultural innovation (Larmuseau et al.,

2016). Because of this ongoing discussion, research should test ovulatory shifts in other cultures too (Henrich et al., 2010).

Suggestions for planning future and reading past cycle studies

The two most interesting takeaways from our researcher degrees of freedom simulations (see supportive website, **osf.io/pbef2**) might be that a) optional stopping and outcome switching had worse impacts than random covariates or switching between narrow, broad, and continuous fertile window estimates, and that b) false positives were acceptably rare (less than 5% in most conditions) if one simply applies a significance threshold of .01. The latter result only holds if researchers behaved as simulated and really stopped at $p < .05$ (Nelson, Simmons, & Simonsohn, 2016), but might provide a useful guide to reading the older, non-preregistered literature.

Although it is difficult to compute an equivalent of Cohen's *d* for multilevel models, our comparable effect size estimates ranged from 0.12 to 0.43. These effect sizes are disattenuated for measurement error in the predictor, but not in the outcome. Some were hence only a quarter of the smallest effect size (0.4) considered in Gangestad et al.'s (2016) simulations and sample size recommendations. Empirically, had we used sample sizes like the studies we were replicating, none of the effects reported here would have been significant. Whether the fertility predictor was formed based on forward- or backward-counting, narrow, broad, or continuous fertile phases seemed to make less of a difference (Figure 4), except that predictors using more data are preferable and that (pre-)menstruation should be adjusted for. While the absolute sizes of the effects we found were not huge, their practical implications might still be noteworthy. The effects on in-pair desire are, for instance, comparable with reported effects of hormonal contraceptive use on sexual desire in a randomised controlled trial (Zethraeus et al., 2016). Moreover, we found evidence for substantial inter-individual variation, so that effects that are small on average might be substantial for some women.

To fully understand the accompanying cyclical changes going along with ovulation, researchers should collect data over many days per woman (Haselton & Gangestad, 2006; Roney & Simmons, 2013, 2016). We have released our study code to make it easier to conduct online diary studies like this one using **formr.org** (Arslan & Tata, 2016). We have also released our data cleaning code, and our code for computing menstrual onsets as potential groundwork for a standard operating procedure. We welcome improvements to this procedure that can be publicly shared. Despite their suboptimal reliability we think using day counting methods is justified by the much larger amount of data that can be and have already been collected efficiently (e.g., in the numerous cycle tracking apps). Still, to directly test mechanisms, hormonal assays, especially repeated ones (Jones et al., 2017, 2018; Roney & Simmons, 2013), are needed as converging evidence. Potentially these two designs can be fruitfully merged (Roney & Simmons, 2013), so that fertile window estimates are used to multiply

impute hormonal assays in a planned missing design. Future research should not only examine and compare the noise (unreliability), but also potential bias engendered by counting and other methods (e.g. underestimates for women with short cycles). Future researchers would improve their odds of detecting an effect by improving the reliability of change for outcomes and predictors, collecting data on more women, more days, or ideally by doing all of this.

Although we fail to conceive any reasonable non-hormonal or non-causal alternative explanations for the changes we observe mid-cycle, these inferences could be strengthened through a true randomised control group. We suggest that future hormonal contraceptive RCTs collect diary data across several full cycles in both experimental groups before and after randomisation. By doing so, we would be able to assess differences caused by contraceptive pills across the whole cycle, not just in e.g. the luteal phase, and we would have sufficiently reliable within-subject data to examine heterogeneity in the response to contraceptive pills. Future studies should also attempt to better test whether awareness of being in the fertile window drives any effects.

Conclusions

In a high-powered, preregistered, within-subject diary study, we replicated main effects of ovulatory increases in self-perceived desirability, as well as extra-pair and in-pair sexual desire and behaviour. We failed to replicate reported ovulatory increases in partner mate retention behaviour and clothing style, and found only ambiguous support for increases in sexual behaviour. In contrast to previous reports, we found no evidence that sexual desire shifted more strongly among women who deemed their partner less sexually attractive. Previous studies often had inadequate power, sometimes used suboptimal between-subject designs, and none were preregistered. Hence, several previous reports of ovulatory shifts and moderators thereof may have been false positives. We do not rule out changes along other dimensions or moderators that we and others have not tested, but large, well-designed, preregistered studies will be necessary to show these credibly. Alternatively, our data are consistent with the theory that ovulatory increases reflect generalized changes in sexual motivation, serving the adaptive function to avoid costs associated with sex when it will not lead to conception (Roney & Simmons, 2013, 2016). Further work should directly test competing theories against each other.

References

Alvergne, A., & Lummaa, V. (2010). Does the contraceptive pill alter mate choice in humans? *Trends in Ecology & Evolution*, *25*, 171–179. doi:10.1016/j.tree.2009.08.003

Arslan, R. C. (2018). Using 26 thousand diary entries to show ovulatory changes in sexual desire and behaviour: Supportive website https://rubenarslan.github.io/ovulatory_shifts/. *Zenodo*. doi:10.5281/zenodo.1243038

Arslan, R. C., & Penke, L. (2015). Evolutionary Genetics. In *The Handbook of Evolutionary Psychology* (Vol. 2, pp. 1047–1066). New York: Wiley. doi:10.1002/9781119125563.evpsych245

Arslan, R. C., & Tata, C. (2016). *formr.org: v0.12.6*. doi:10.5281/zenodo.60957

Asendorpf, J. B., & Wilpers, S. (1998). Personality effects on social relationships. *Journal of Personality and Social Psychology*, *74*, 1531–1544. doi:10.1037/0022-3514.74.6.1531

Back, M. D., Küfner, A. C. P., Dufner, M., Gerlach, T. M., Rauthmann, J. F., & Denissen, J. J. A. (2013). Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. *Journal of Personality and Social Psychology*, *105*, 1013–1037. doi:10.1037/a0034431

Bafumi, J., & Gelman, A. (2006). Fitting multilevel models when predictors and group effects correlate. *Social Science Research Network*. doi:10.2139/ssrn.1010095

Baird, D. D., McConnaughey, D. R., Weinberg, C. R., Musey, P. I., Collins, D. C., Kesner, J. S., … Wilcox, A. J. (1995). Application of a method for estimating day of ovulation using urinary estrogen and progesterone metabolites. *Epidemiology (Cambridge, Mass.)*, *6*, 547–550.

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. doi:10.1016/j.jml.2012.11.001

Barta, W. D., Tennen, H., & Litt, M. D. (2012). Measurement reactivity in diary research. In M. R. Mehl & M. Connor (Eds.), *Handbook of research methods for studying daily life* (pp. 108–123). New York, NY: Guilford Press.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv:1506.04967 [stat]*. Retrieved from http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv:1406.5823 [stat]*. doi:arXiv:1406.5823

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Blake, K. R., Dixson, B. J. W., O'Dean, S. M., & Denson, T. F. (2016). Standardized protocols for characterizing women's fertility: A data-driven approach. *Hormones and Behavior*, *81*, 74–83. doi:10.1016/j.yhbeh.2016.03.004

Brewis, A., & Meyer, M. (2005). Demographic evidence that human ovulation is undetectable (at least in pair bonds). *Current Anthropology*, *46*, 465–471. doi:10.1086/430016

Bürkner, P.-C. (in press). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2015). Stan: A probabilistic programming language. *Journal of Statistical Software*.

Caruso, S., Agnello, C., Malandrino, C., Lo Presti, L., Cicero, C., & Cianci, S. (2014). Do hormones influence women's sex? Sexual activity over the menstrual cycle. *The Journal of Sexual Medicine*, *11*, 211–221. doi:10.1111/jsm.12348

Deschner, T., Heistermann, M., Hodges, K., & Boesch, C. (2003). Timing and probability of ovulation in relation to sex skin swelling in wild West African chimpanzees, Pan troglodytes verus. *Animal Behaviour*, *66*, 551–560. doi:10.1006/anbe.2003.2210

Dixson, A. F. (2012). *Primate sexuality: Comparative studies of the prosimians, monkeys, apes, and humans 2nd edition*. Oxford: Oxford University Press.

Durante, K. M., Li, N. P., & Haselton, M. G. (2008). Changes in women's choice of dress across the ovulatory cycle: Naturalistic and laboratory task-based evidence. *Personality & Social Psychology Bulletin*, *34*, 1451–1460. doi:10.1177/0146167208323103

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904. doi:10.1007/s11192-011-0494-7

Fehring, R. J., Schneider, M., & Raviele, K. (2006). Variability in the phases of the menstrual cycle. *Journal of Obstetric, Gynecologic*, *35*, 376–384. doi:10.1111/j.1552-6909.2006.00051.x

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*, 120–128. doi:10.1037/a0024445

Fincham, F. D., & May, R. W. (2017). Infidelity in romantic relationships. *Current Opinion in Psychology*, *13*, 70–74. doi:10.1016/j.copsyc.2016.03.008

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*, 1011–1027.

Fleeson, W. (2004). Moving personality beyond the person-situation debate: The challenge and the opportunity of within-person variability. *Current Directions in Psychological Science*, *13*, 83–87. doi:10.1111/j.0963-7214.2004.00280.x

Fleischman, D. S., & Fessler, D. M. (2007). Differences in dietary intake as a function of sexual activity and hormonal contraception. *Evolutionary Psychology*, *5*, 642–652.

Gangestad, S. W. (2016). Comment: Wood et al.'s (2014) speculations of inappropriate research practices in ovulatory cycle studies. *Emotion Review*, *8*, 87–90. doi:10.1177/1754073915580400

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. doi:10.1016/j.evolhumbehav.2014.02.012

Gangestad, S. W., Haselton, M. G., Welling, L. L. M., Gildersleeve, K., Pillsworth, E. G., Burriss, R. P., … 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. doi:10.1016/j.evolhumbehav.2015.09.001

Gangestad, S. W., & Simpson, J. A. (2000). The evolution of human mating: Trade-offs and strategic pluralism. *Behavioral and Brain Sciences*, *23*, 573–587. doi:10.1017/S0140525X0000337X

Gangestad, S. W., & Thornhill, R. (2008). Human oestrus. *Proceedings of the Royal Society B: Biological Sciences*, *275*, 991–1000. doi:10.1098/rspb.2007.1425

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*, 975–982. doi: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 B: Biological Sciences*, *272*, 2023–2027. doi:10.1098/rspb.2005.3112

Gangestad, S. W., Thornhill, R., & Garver-Apgar, C. E. (2015). Women's sexual interests across the ovulatory cycle. *The Handbook of Evolutionary Psychology*. doi:10.1002/9781119125563.evpsych114

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*, 460.

Gerlach, T. M., Arslan, R. C., Schultze, T., Reinhard, S. K., & Penke, L. (in press). Predictive validity and adjustment of ideal partner preferences across the transition into romantic relationships. *Journal of Personality and Social Psychology*.

Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014a). Do women's mate preferences change across the ovulatory cycle? A meta-analytic review. *Psychological Bulletin*, *140*, 1205–1259. doi:10.1037/a0035438

Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014b). Meta-analyses and p-curves support robust cycle shifts in women's mate preferences: Reply to Wood and Carden (2014) and Harris, Pashler, and Mickes (2014). *Psychological Bulletin*, *140*, 1272–1280. doi:10.1037/a0037714

Grotzinger, A. D., Mann, F. D., Patterson, M. W., Herzhoff, K., Tackett, J. L., Tucker-Drob, E. M., & Paige Harden, K. (2017). Twin models of environmental and genetic influences on pubertal development, salivary testosterone, and estradiol in adolescence. *Clinical Endocrinology*. doi:10.1111/cen.13522

Harris, C. R., Pashler, H., & Mickes, L. (2014). Elastic analysis procedures: An incurable (but preventable) problem in the fertility effect literature. Comment on Gildersleeve, Haselton, and Fales (2014). *Psychological Bulletin*, *140*, 1260–1264. doi:10.1037/a0036478

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. doi:10.1016/j.yhbeh.2005.10.006

Haselton, M. G., & Gildersleeve, K. (2016). Human ovulation cues. *Current Opinion in Psychology*, *7*, 120–125. doi:10.1016/j.copsyc.2015.08.020

Haselton, M. G., Mortezaie, M., Pillsworth, E. G., Bleske-Rechek, A., & Frederick, D. A. (2007). Ovulatory shifts in human female ornamentation: Near ovulation, women dress to impress. *Hormones and Behavior*, *51*, 40–45. doi:10.1016/j.yhbeh.2006.07.007

Havlíček, J., Cobey, K. D., Barrett, L., Klapilová, K., & Roberts, S. C. (2015). The spandrels of Santa Barbara? A new perspective on the peri-ovulation paradigm. *Behavioral Ecology*, *26*, 1249–1260. doi:10.1093/beheco/arv064

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences*, *33*, 111–135. doi:10.1017/S0140525X10000725

Hill, S. E., & Durante, K. M. (2009). Do Women Feel Worse to Look Their Best? Testing the Relationship Between Self-Esteem and Fertility Status Across the Menstrual Cycle. *Personality and Social Psychology Bulletin*, *35*, 1592–1601. doi:10.1177/0146167209346303

Inzlicht, M., Gervais, W., & Berkman, E. (2015). Bias-correction techniques alone cannot determine whether ego depletion is different from zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015. *Social Science Research Network*. doi:10.2139/ssrn.2659409

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

Jones, B. C., Hahn, A. C., Fisher, C. I., Wang, H., Kandrik, M., & DeBruine, L. M. (2018). General sexual desire, but not desire for uncommitted sexual relationships, tracks changes in women's hormonal status. *Psychoneuroendocrinology*, *88*, 153–157. doi:10.1016/j.psyneuen.2017.12.015

Jones, B. C., Hahn, A. C., Fisher, C. I., Wang, H., Kandrik, M., Han, C., … DeBruine, L. M. (2017). Within-woman hormone-attractiveness correlations are not simply byproducts of between-women hormone-attractiveness correlations. *bioRxiv*, 136515. doi:10.1101/136515

Jones, B. C., Hahn, A. C., Fisher, C. I., Wang, H., Kandrik, M., Han, C., … DeBruine, L. M. (2017). Women's preferences for facial masculinity are not related to their hormonal status. *bioRxiv*, 136549. doi:10.1101/136549

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist*, *43*, 23–34. doi:10.1037/0003-066X.43.1.23

Landolt, M. A., Lalumière, M. L., & Quinsey, V. L. (1995). Sex differences in intra-sex variations in human mating tactics: An evolutionary approach. *Ethology and Sociobiology*, *16*, 3–23. doi:10.1016/0162-3095(94)00012-V

Lang, F. R., Lüdtke, O., & Asendorpf, J. B. (2001). Testgüte und psychometrische Äquivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittelalten und alten Erwachsenen. *Diagnostica*, *47*, 111–121. doi:10.1026//0012-1924.47.3.111

Larmuseau, M. H., Matthijs, K., & Wenseleers, T. (2016). Cuckolded fathers rare in human populations. *Trends in Ecology & Evolution*, *31*, 327–329. doi:10.1016/j.tree.2016.03.004

Larson, C. M., Haselton, M. G., Gildersleeve, K. A., & Pillsworth, E. G. (2013). Changes in women's feelings about their romantic relationships across the ovulatory cycle. *Hormones and Behavior*, *63*, 128–135. doi:10.1016/j.yhbeh.2012.10.005

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2016, January 14). Ambitious p-hacking and p-curve 4.0. *Data Colada*. Retrieved from http://datacolada.org/45

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in r. *Journal of Statistical Software*, *74*, 1–26. doi:10.18637/jss.v074.i11

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. doi:10.1126/science.aac4716

Pillsworth, E. G., & Haselton, M. G. (2006a). Women's sexual strategies: The evolution of long-term bonds and extrapair sex. *Annual Review of Sex Research*, *17*, 59–100. doi:10.1080/10532528.2006.10559837

Pillsworth, E. G., & Haselton, M. G. (2006b). Male sexual attractiveness predicts differential ovulatory shifts in female extra-pair attraction and male mate retention. *Evolution and Human Behavior*, *27*, 247–258. doi:10.1016/j.evolhumbehav.2005.10.002

Pillsworth, E. G., Haselton, M. G., & Buss, D. M. (2004). Ovulatory shifts in female sexual desire. *Journal of Sex Research*, *41*, 55–65.

Ranehill, E., Zethraeus, N., Blomberg, L., von Schoultz, B., Hirschberg, A. L., Johannesson, M., & Dreber, A. (2017). Hormonal Contraceptives Do Not Impact Economic Preferences: Evidence from a Randomized Trial. *Management Science*. doi:10.1287/mnsc.2017.2844

Revelle, W. (2017). *psych: Procedures for psychological, psychometric, and personality research*. Retrieved from https://CRAN.R-project.org/package=psych

Roberts, B. W., & Caspi, A. (2001). Personality development and the person-situation debate: It's déjà vu all over again. *Psychological Inquiry*, *12*, 104–109. doi:10.1207/S15327965PLI1202_04

Roney, J. R. (2009). The role of sex hormones in the initiation of human mating relationships. In P. T. Ellison & P. B. Gray (Eds.), *The Endocrinology of Social Relationships* (pp. 246–269). Cambridge: Harvard University Press.

Roney, J. R., & Simmons, Z. L. (2013). Hormonal predictors of sexual motivation in natural menstrual cycles. *Hormones and Behavior*, *63*, 636–645. doi:10.1016/j.yhbeh.2013.02.013

Roney, J. R., & Simmons, Z. L. (2016). Within-cycle fluctuations in progesterone negatively predict changes in both in-pair and extra-pair desire among partnered women. *Hormones and Behavior*, *81*, 45–52. doi:10.1016/j.yhbeh.2016.03.008

Roney, J. R., & Simmons, Z. L. (2017). Ovarian hormone fluctuations predict within-cycle shifts in women's food intake. *Hormones and Behavior*, *90*, 8–14. doi:10.1016/j.yhbeh.2017.01.009

Scheib, J. E., Gangestad, S. W., & Thornhill, R. (1999). Facial attractiveness, symmetry and cues of good genes. *Proceedings of the Royal Society B: Biological Sciences*, *266*, 1913–1917. doi:10.1098/rspb.1999.0866

Schilling, K. M., Straus, H., Arslan, R. C., Gerlach, T. M., & Penke, L. (2014). Cycle shifts and pill dosage effects. doi:10.17605/osf.io/98w3h

Schwarz, S., & Hassebrauck, M. (2008). Self-perceived and observed variations in women's attractiveness throughout the menstrual cycle—a diary study. *Evolution and Human Behavior*, *29*, 282–288. doi:10.1016/j.evolhumbehav.2008.02.003

Shrout, P., & Lane, S. P. (2012). Psychometrics. In *Handbook of research methods for studying daily life*. Guilford Press.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*, 160384. doi:10.1098/rsos.160384

Stirnemann, J. J., Samson, A., Bernard, J.-P., & Thalabard, J.-C. (2013). Day-specific probabilities of conception in fertile cycles resulting in spontaneous pregnancies. *Human Reproduction*, *28*, 1110–1116. doi:10.1093/humrep/des449

van Elk, M., Matzke, D., Gronau, Q., Guang, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology*, *6*, 1365. doi:10.3389/fpsyg.2015.01365

Wegienka, G., & Baird, D. D. (2005). A comparison of recalled date of last menstrual period with prospectively recorded dates. *Journal of Women's Health*, *14*, 248–252. doi:10.1089/jwh.2005.14.248

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *22*, 392–399. doi:10.2307/2346786

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*, 95–114. doi:10.1111/1467-9868.00374

Wood, W., Kressel, L., Joshi, P. D., & Louie, B. (2014). Meta-analysis of menstrual cycle effects on women's mate preferences. *Emotion Review*. doi:10.1177/1754073914523073

Zethraeus, N., Dreber, A., Ranehill, E., Blomberg, L., Labrie, F., Schoultz, B. von, … Hirschberg, A. L. (2017). A first-choice combined oral contraceptive influences general well-being in healthy women: A double-blind, randomized, placebo-controlled trial. *Fertility and Sterility*, *107*, 1238–1245. doi:10.1016/j.fertnstert.2017.02.120

Zethraeus, N., Dreber, A., Ranehill, E., Blomberg, L., Labrie, F., von Schoultz, B., … Hirschberg, A. L. (2016). Combined oral contraceptives and sexual function in women - a double-blind, randomized, placebo-controlled trial. *The Journal of Clinical Endocrinology and Metabolism*, *101*, 4046–4053. doi:10.1210/jc.2016-2032

Zietsch, B. P., Lee, A. J., Sherlock, J. M., & Jern, P. (2015). Variation in women's preferences regarding male facial masculinity is better explained by genetic differences than by previously identified context-dependent effects. *Psychological Science*, *26*, 1440–1448. doi:10.1177/0956797615591770

Supplement

Supplementary Table 1. Descriptions of robustness checks.

Model	Description	Exclusion criteria
M_1	Baseline robustness check model. A random intercept for the participant, fixed effects for hormonal contraception, (pre-)menstruation, fertility (backward-counted from the observed *or* inferred next menstrual onset), and average fertility (to ensure within-subject estimates). In Wilkinson notation: outcome ~ (fertile_window_probability + premenstrual_phase + menstruation) * hormonal_contraceptive_user + average_fertile_window_probability+ (1 l person)	Minimal ("all")[1]
M_r1	M_1 + Allowed individual differences in fertility effect: a random slope for the fertile window probability, the premenstruation dummy, and the menstruation dummy.	
M_e2	M_1 + first set of preregistered exclusion criteria.	lax[1]
M_e3	M_e2 + amended set of preregistered exclusion criteria except for two criteria.	conservative[1]
M_e4	M_e3 + excluding those who felt stressed and those who said they might have irregular cycles, but were very unsure.	strict[1]
M_e5	M_1 + excluded 1251 diary days (4% of all) where participants a) gave the same answer to all Likert items (n=23), b) accessed the diary later or earlier than intended due to technical problems (n=896), or c) took more than 24 hours (n=376) or less than a minute (n=30) to finish filling out the diary	Without potentially unreliable data.
M_e6	M_1 + excluding women who were trying to get pregnant.	No women who were trying to get pregnant.
M_p1	M_1 + FP was continuous and based on *confirmed* next menstrual onsets, without onsets *inferred* from average cycle length.	
M_p2	M_1 + FP was continuous and based on *last* menstrual onsets, i.e. forward-counting.	

Model	Description	Exclusion criteria
M_p3	M_p1, but FP restricted to a broad window.[2]	
M_p4	M_p1, but FP restricted to a narrow window.[2]	
M_p5	M_p2, but FP restricted to a broad window.[2]	
M_p6	M_p2, but FP restricted to a narrow window.[2]	
M_p7	M_1 + FP was continuous and based on *inferred* next onsets based on reported average cycle length.	
M_p8	M_1 + Restricted the analysis to women within the range in which most previous fertility estimates were given	No women with average cycle lengths outside 20-40
M_c1	M_1 + Adjusted for self-esteem.	
M_c2	M_1 + No adjustment for average fertile window probability.	
M_c3	M_c2 + No adjustment for (pre-)menstruation dummies	
M_c4	M_1 + Adjusted for weekday and number of weeks since starting the diary.	
M_c5	M_1 + Adjusted for time of response and time taken for response (log10+1).	
M_c6	M_1 + Modelled autocorrelation of order 1.	
M_c7	M_1 + Modelled moving averages of order 1.	
M_c8	M_p3, but with adjustments for (pre-)menstruation and average fertility, often did not converge see [1]	
M_c9	M_1 + adjusted for thin-plate splines (nonlinear effects) for the number of days since the diary began (one variable for days filled out and one including missing days), separate splines for hormonal contraceptive (non-)users	Only people with more than 37 days filled out.
M_d1	M_p2 + Between-subject design.	Took only the first day of the diary for every participant.

Model	Description	Exclusion criteria
M_d2	M_p6 + Within-subject design with two days per participant.	Only one high- and one low-fertility day per participant.
M_d3	M_p6 + Within-subject design with four days per participant.	Only two high- and two low-fertility day per participant.
M_d4	M_1 + No observed cycle lengths shorter than 20 days.	If time between menstrual onsets was lower than 20 days, excluded.
M_d5	M_p3 + Within-subject design with high- and low-fertility averaged separately per participant (so ignoring varying number of days per participant).	All days outside the window.
M_m1	M_1 + Instead of letting hormonal contraception status moderate the fertility and menstruation predictors, we differentiated by: hormonal, fertility awareness, barrier/abstinence, none	Excluding women who use other methods than these (e.g. partner sterilisation).
M_m2	M_1 + Age group (18-20, 20-25, 25-30, 30-35, 35 and older) moderates the FP.	
M_m3	M_1 + A weekend dummy moderates the FP.	
M_m4	M_1 + Weekday dummies moderate the FP.	
M_m5	M_1 + Maximal applicable exclusion threshold (all, lax, conservative, strict) moderates the FP.	
M_m6	M_1 + Cycle length (19-25, 25-30, 30-35, 35-40) moderates the FP.	
M_m7	M_1 + Self-reported certainty about menstruation regularity/cycle length moderates the FP.	
M_m8	M_1 + Self-reported menstruation regularity moderates the FP.	
M_m9	M_1 + Cohabitation status (same apartment, same city, long-distance) moderates the FP.	
M_m10	M_1 + Relationship status (partnered, engaged, married) moderates the FP.	

Notes. These are the robustness checks that we conducted for all outcomes. FP = Fertility predictor. [1] For definitions , see Figure 1. [2] Using windowed predictors, the effects of menstruation and average fertility could no longer be stably estimated because of the reduced number of days, so these adjustments were omitted. The complete robustness analyses, including all code and results can be found on the supportive website at **https://rubenarslan.github.io/ovulatory_shifts/3_fertility_robustness.html** and an extended version of this table can be found at **https://rubenarslan.github.io/ovulatory_shifts/3_robustness_checks_table.html**