Title: Hormonal Contraceptive Use and Women's Sexuality and Well-Being: Estimating Treatment Effects and Their Heterogeneity Based on Longitudinal Data

July 12th, 2023

Dear Prof. Dr. Evans, dear reviewers,

Thank you very much for the helpful and constructive feedback on our programmatic registered report stage 1 on heterogeneous effects of contraception and for the opportunity to revise and resubmit the manuscript to *Peer Community In - Registered Reports*.

We were happy to hear that you and two reviewers found our study of interest. We greatly appreciate yours and the reviewers' insightful comments and valuable feedback on ways to further improve the manuscript. Thus, we have carefully considered and responded to all points raised by you and the reviewers.

We have improved the manuscript based on yours and the reviewers' comments in the following areas: We clarified our theoretical considerations (addressing comments R2.1., R2.2, and R2.3) and how we will interpret results based on these considerations (addressing comment R1.6.). We provided more information about the included variables (addressing comments R1.1.a, R1.10., R2.9., R2.10., and R2.11). By adding a new Table 2 outlining the exclusion criteria and reasons for exclusions we aim to address comments by the reviewers concerning the understandability of our anticipated sample based on the manuscript (addressing comments R1.1., R1.2., R1.2.a, R1.4., R2.7., R2.8., and R2.9.). Besides giving a more precise description in the manuscript about our planned analyses we additionally improved our planned statistical analyses based on the reviewers' feedback (addressing comments R1.3., R1.5., R1.8., R1.11., R2.4., R2.5., R2.12., and R2.13.). Even though we appreciated the many interesting ideas and additional research questions proposed by the reviewers, we decided against including many interesting but complex additional analyses to prevent overloading the manuscript (addressing comments R1.1.a, R1.1.b, and R1.7.). Finally, we provided more information about the simulation (addressing comment R1.9).

Throughout the manuscript we made minor changes to improve readability and to ensure consistency throughout the manuscript. In addition, we rephrased the title to "Hormonal Contraceptive Use and Women's Sexuality and Well-Being: Estimating Treatment Effects and Their Heterogeneity Based on Longitudinal Data".

We uploaded two versions of the manuscript, one in which all changes from the initially submitted version of the manuscript are presented in blue font to ease the review process. We think that these changes have further strengthened our manuscript and would be grateful if you consider it for stage 1 acceptance. We responded to each suggestion below and numbered comments to make them easier to identify.

Best regards,

Laura Botzet, on behalf of all co-authors

Recommender's remarks:

Thank-you again for submitting your Stage 1 to PCI:RR. I thoroughly enjoyed reading through your work and the prepared materials shared via the OSF page. I wish to make it immediately explicit in framing my evaluation of the manuscript that I do not have experience researching contraceptive use. However, I sincerely hope that my methodological/open scholarship expertise and contributions to publishing 11 (and counting!) of my own RRs provides a rigorous basis for providing you with fruitful feedback. My goal is to help you make this work the best it can be and I look forward to supporting you on that basis.

Thank you so much. We really appreciate your support and very valuable and helpful feedback.

Just a few hours ago the second review was submitted so I am pleased to confirm that I now have two appropriate reviews and I would like to encourage you to make revisions considering their supportive and constructive feedback as seen below. Their comments and suggestions cover a range of themes from the need to justify your analysis approach further to consideration of other factors and models. The crux of their feedback, and of this paper as a whole, is the modelling (both conceptual and statistical) where there is opportunity to go in many directions and subsequently where you may wish to further justify or reconsider decisions previously made in context of the available options the data allows. I have no interest in dictating to you which of these need full and comprehensive implementation and which do not, but I encourage you to be systematic and open in considering all their feedback, and providing a clear response to each comment regardless of whether they are actioned. I particularly appreciate use of 'tracked changes' or similar approaches to log changes, but please ensure your response to the feedback is as comprehensive and accessible as possible to facilitate the next stage of review.

We have provided answers to all comments and addressed all proposed changes below. To facilitate the review process we numbered comments and copied parts of the manuscript into our revision letter. In addition, we uploaded two versions of the manuscript, one in which all changes from the initially submitted version of the manuscript are presented in blue font to ease the review process.

From my own reading, your work is a really interesting analysis of the substantive PAIRFAM dataset and considers two sets of highly important outcomes. I particularly appreciated the critical tone adopted when considering the conclusions and limitations of the extant literature. Your introduction provides a clear overview of the limitations to our current approach to studying these effects and whilst you identified a range of dimensions which problematise our current understanding well, both reviewers suggested further justification for some of your core decisions which would help to build a more convincing and rigorous basis for your study. You could provide a little more detail on the simulation and access to the data provided by Tita Gonzalez Avilés too, to make it clearer as to how you used this information. More broadly,

throughout there were occasionally some sentences which were a little wordy or heavy to follow and I encourage you to proof-read through the manuscript with this in mind to ensure it is as accessible and clear as possible, particularly as the reviewers highlighted some specific areas where this obfuscated the meaning of the content. Whilst I would have liked a little more annotation alongside the code, and I encourage you to revisit this as you work through the comments above, all the materials were well-structured and easily accessed.

We now provide more justification for our core decisions (see response to R1.6.) and more detail on the simulation (see response to R1.9). In addition, we proof-read the manuscript with a focus on making the manuscript more accessible and as clear as possible. Alongside the updated code we now provide additional annotations alongside the code and explain our planned analyses in more detail in the manuscript (see responses to R1.3., R.1.5., R1.7., R2.5., R2.6., R2.12.).

In sum, I would like to encourage you to reflect upon the feedback provided, and to resubmit when you have systematically reconsidered the core decisions made behind the models proposed. I sincerely hope you find this feedback and process to be fruitful in supporting the further clarity and impact of the work, and I look forward to hearing from you in due course,

Stay safe and take care,

Dr Thomas Rhys Evans (Tom/He/Him)

Reviewer #1's remarks:

The authors propose a series of analyses to be conducted on a large, longitudinal dataset containing data from 6,565 women collected over the course of 13 waves (waves collected annually, average of 6 waves per participant). The authors seek to estimate the effect of hormonal contraceptive (HC) use on both sexual outcomes and well-being. Further, they also seek to investigate the heterogeneity in women's responses to HC treatment and if women's treatment responses predict their future HC use. Generally, I found the research questions to be well supported and the relevant literature to be well described. The research questions the authors propose are of keen scientific interest, and the authors' choice of study design is appropriate to answer these questions. That is, naturalistic, longitudinal studies are certainly one vital piece of the puzzle needed to understand how HC use impacts both women's sexuality and well-being.

While I am not as statistically literate as the authors, their statistical approach appears well conceived and appropriate, although given my limited knowledge of Bayesian models in practice, I will ask a few clarifying questions below to ensure that these approaches are indeed appropriate and would not benefit from supplementation with additional exploratory analyses or alternate approaches.

We really appreciate the clarifying questions as they highlight places in the manuscript where we were not clear enough about our statistical approach. We hope that our answers to these questions and the changes we made in the manuscript address all concerns (see responses to R1.3., R1.7., R1.8., R1.9., and R1.11)

While of acceptable standards, there are a few points in the methodology and data analysis plan which I believe could use additional clarification before I recommend this Stage 1 registered report, which I detail below. My only major concerns with the data analytic strategy lies in the authors' abilities to answer the questions they seek to answer given the limited specificity of the data available and if this is being appropriately modeled, given the lack of a statistical model included in the report. Finally, I would like to see the authors provide what their interpretation will be for potential results, to get an idea of what the data analysis will reveal. Overall, I found this to be an impressively well written registered report and an interesting research study.

We thank the reviewer very much for this summary and for the very helpful specific points addressed in the following. We have added additional clarification about the methodology and the data analysis plan, especially concerning exclusion criteria and robustness analyses (see responses to R1.1., R1.1.a, R1.1b, R1.2., R1.2a, R1.4., R1.10.). The new version of this manuscript includes notation for all statistical models (see responses to R1.3. and R1.5.). In addition, we provided more information about potential interpretations, especially in the case when different analysis plans reach different conclusions (see response to R1.6.). Please see below for more detail how we addressed each of the specific points raised by the reviewer.

Specific points:

R1.1. The variable options available in this dataset are somewhat limited. For example, there is an option for IUDs in type of contraceptive used, but no specification of if this is a hormonal or nonhormonal IUD – is this information provided in a follow up question? How will the authors determine if a woman is using HCs if she reports an IUD? I saw that IUD was a potential exclusion criterion, but that does prevent analysis of changing methods to/from HCs in the event that one of the options is an IUD (hormonal or copper). How do the authors plan to handle this confound? This seems important, given the popularity of IUDs as both a hormonal and non-hormonal contraceptive option.

We thank the reviewer for this comment. Unfortunately, participants were not asked whether they used a hormonal or copper intrauterine device (sadly, this is often the case in panel datasets). For the main analyses we decided to code the choice *intrauterine device* as hormonal if participants had indicated earlier in the survey that they used *other hormonal method (implant, patch, ring)*, assuming that women who use a hormonal intrauterine device would classify this as another hormonal method after the option *birth control pill, mini pill*. If participants only indicated that they use an *intrauterine device* but no hormonal method, this was coded as non-hormonal contraception for the main analyses.

As the transformation of the variable hormonal contraception was also an issue in comments R1.1.a and R2.9., we describe the transformation of this variable in more detail in the section *Methods - Variables*:

The predictor hormonal contraception will be based on the items about the contraceptive method; participants were able to report multiple contraceptive methods. Hormonal contraception will be coded as 0 if participants indicated that they use no contraceptive method at all. The variable hormonal contraception will also be coded as 0 if participants indicated that they use no hormonal contraceptive method and at least one of the following methods: condom*;* intrauterine device[4]*;* diaphragm, foam, suppository, gel*;* natural birth control*;* female sterilization*;* male sterilization*; or* withdrawal method, coitus interruptus*. In addition, the variable hormonal contraception will be coded as 0 if participants were never sexually active in their life, as these participants were not asked about their contraceptive method. The variable hormonal contraception will be coded as 1 if participants indicated that they use* a birth control pill*,* mini pill *or* other hormonal method (implant, patch, ring) *even if they additionally use non-hormonal methods. Exclusion criteria for main as well as robustness analyses based on the contraceptive method are described above.*

In addition, we added a footnote to the choice *intrauterine device* explaining our decision outlined above:

Participants were not asked whether they used a hormonal or copper intrauterine device. Therefore, we will code the choice intrauterine device as hormonal if participants had indicated earlier in the survey that they use other hormonal method (implant, patch, ring)*, assuming that women who use a hormonal intrauterine device would classify this as another hormonal method after the option* birth control pill, mini pill*. If participants only indicated that they use an* intrauterine device *but no hormonal method, this was coded as non-hormonal contraception.*

We decided to include intrauterine device users in our main analyses, even though the coding as hormonal vs. non-hormonal might be wrong in some cases. As the reviewer notes, hormonal as well as copper intrauterine devices are a popular contraceptive method and we wanted to include them in our analyses. Nevertheless, as noted in the section *Methods - Exclusion Process and Participants*, we will perform an additional robustness analysis excluding waves in which women indicated using an intrauterine device, to make sure that potential effects are not due to a wrong classification of intrauterine devices as hormonal or non-hormonal.

The lack of a clear explanation of our exclusion criteria for our main and robustness analyses and specific reasons for these criteria were also criticized in comments R1.2, R1.2.a, R2.7, R2.8, and R2.9. Therefore, we rewrote the section *Methods - Exclusion Process and Participants* and added Table 2 summarizing exclusion criteria, reasons for exclusion, and unit(s) that will be excluded.

We plan to exclude individuals who did not identify as female. Furthermore, once a woman crossed the age of 50 or reported to be (post-)menopausal, her data (including subsequent waves) will be excluded, but previous waves of data collection will remain in the analysis. In addition, we plan to exclude all individual waves of data in which participants indicated being in a homosexual relationship or only reported homosexual relationships in the past, were pregnant, trying to become pregnant, gave birth to a child in the last year, were currently breastfeeding, or indicated using the morning-after-pill, or an unknown contraceptive method.

In further separate robustness analyses we plan to additionally exclude waves in which participants indicated that they are sterilized, as well as all subsequent waves of those participants.We will also exclude all waves in which participants indicated that their partner is sterilized and all waves in which women indicated using no contraceptive method[3], an intrauterine device as a contraceptive method, or hormonal methods other than the oral contraceptive pill. In addition, we will exclude all waves in which women indicated that they had never been sexually active. All exclusion criteria, reasons for exclusion, and excluded unit(s) are summarized in Table 2.

Table 2

Exclusion criteria, reasons for exclusion, and excluded units

Main analyses		
Exclusion criteria	**Reasons for exclusion**	**Excluded unit(s)**
identifying as non-female	potential hormonal influences	current and all subsequent waves
older than 50 years	potential hormonal influences	current and all subsequent waves
(post-)menopausal	potential hormonal influences	current and all subsequent waves
only homosexual relationships	no need to use contraceptives to prevent pregnancy	current wave
pregnant	potential hormonal influences	current wave
trying to become pregnant	no need to use contraceptives to prevent pregnancy	current wave
gave birth in the last year	potential hormonal influences	current wave
breastfeeding	potential hormonal influences	current wave
using the morning-after-pill as a contraceptive method	potential hormonal influences	current wave
using an unknown contraceptive method	not possible to classify method as hormonal or non-hormonal	current wave
Further robustness analyses		
Exclusion criteria	**Reasons for exclusion**	**Excluded unit(s)**
sterilized	no need to use contraceptives to prevent pregnancy	current and all subsequent waves
partner sterilized	no need to use contraceptives to prevent pregnancy	current wave
using no contraceptive method	imprecise classification as non-hormonal in main analyses	current wave
using an intrauterine device as a contraceptive method	imprecise classification as non-hormonal in main analyses	current wave
using other hormonal methods	investigate effects of oral contraceptive pills only	current wave
never sexually active	potentially conditioning on the sexual frequency as an outcome	current and all subsequent waves

R1.1.a Is there any more specificity in the data about what actual type of HC is being used? I worry that given the likelihood that different methods have different effects (and perhaps more meaningfully, different effect sizes) that lumping methods into such broad categories will limit the impact of the results and the reliability of effect size estimates.

Yes, there is more specific data about the contraceptive method as outlined in Table 3 (former Table 2). As reviewer #2 also notes in their comment R2.9. that it was hard to follow how our variable hormonal contraception was built, we describe the transformation of this variable in more detail in the section *Methods - Variables*:

The predictor hormonal contraception will be based on the items about the contraceptive method; participants were able to report multiple contraceptive methods. Hormonal contraception will be coded as 0 if participants indicated that they use no contraceptive method at all. The variable hormonal contraception will also be coded as 0 if participants indicated that they use no hormonal contraceptive method and at least one of the following methods: condom*;* intrauterine device[4]*;* diaphragm, foam, suppository, gel*;* natural birth control*;* female sterilization*;* male sterilization*; or* withdrawal method, coitus interruptus*. In addition, the variable hormonal contraception will be coded as 0 if participants were never sexually active in their life, as these participants were not asked about their contraceptive method. The variable hormonal contraception will be coded as 1 if participants indicated that they use* a birth control pill*,* mini pill *or* other hormonal method (implant, patch, ring) *even if they additionally use non-hormonal methods. Exclusion criteria for main as well as robustness analyses based on the contraceptive method are described above.*

Even though we agree with the reviewer that it would be very interesting to look at effects of specific contraceptive methods, we decided against doing so. We think that we need to answer more broad questions first before we can tackle contraceptive type-specific questions. Therefore, the question whether different methods have different effects is not central for our research question here. The proposed analyses of this registered report are already very complex, and we therefore plan to do things one step at a time. In addition, as we are especially interested in women who switched their contraceptive method, this would result in very complex switching patterns. Therefore, we decided to focus on the broad categories hormonal vs. non-hormonal contraception in this study. As the data and our code are publicly available it would be easy for others (or us) to use our findings as scaffolding to look at more detailed patterns in the future.

R1.1.b Likewise, I am surprised that the authors did not plan to investigate how oral vs non-oral HC methods differed in their effects. I would like to see this question addressed within the current protocol, given that it seems like a decently simple difference to investigate here and would be of interest to other HC researchers. (Perhaps the authors are planning to do this and I misinterpreted this plan)

We agree that the broader oral vs non-oral hormonal contraception contrast is an interesting one. Thus, as outlined in the section *Methods - Exclusion Process and Participants*, we do plan to run additional robustness analyses limiting participants to users of oral contraceptive pills. Here, we want to test the robustness of potential effects of hormonal contraceptive use when limited to a more homogenous hormonal contraceptive type (even though this category still includes combined oral contraceptive pills and mini pills [i.e., progesterone only pills] as we are unfortunately unable to separate those based on the available data). However, we do not plan to run these additional robustness analyses based on a sample limited to non-oral hormonal contraceptive methods, as we expect the sample size to be too small to reach any conclusion given the expected small effect sizes.

R1.2. Instead of excluding women who report homosexual relationships, I would prefer to see models run with and without these exclusions. Additionally, I am a bit confused as to why women using no contraceptive method are being excluded from analysis – it seems to me that they are an important comparison group for HC users, given the goals of the research design. Perhaps the authors had considered this and have another reason for this exclusion, however, I would like to see this explained or an additional exploratory analysis conducted.

The exclusion of homosexual relationships was also criticized by reviewer #2 in comment R2.8. We decided to exclude them because the reasons for using contraceptive methods differ for homosexual women as they do not have to avoid unwanted pregnancies (and will thus probably use hormonal contraceptive methods less often). We think that the decision process which contraceptive method to use will therefore differ notably for homosexual women. This is the reason why we decided to exclude waves in which women were currently in a homosexual relationship or exclusively reported homosexual relationships in the past. We added additional information to the new Table 2 that hopefully clarifies our decision process. Unfortunately, sample sizes for women in non-heterosexual relationships will be too small to reliably reach any conclusion given the expected small effect sizes.

Women using no contraceptive method will be included in the main analyses but excluded in an additional robustness analysis (see section *Methods - Exclusion Process and Participants* for the exclusion process for main and robustness analyses). In PAIRFAM, women who indicated that they had never been sexually active in their life were not asked about their contraceptive method. For the main analysis these women were coded as using no contraceptive method, i.e., a non-hormonal method and we explain this now in more detail in the section *Methods - Variables* (see comments R1.1., R1.1.a, and R.2.9). Of course, this

may introduce some errors as women may use hormonal methods without being sexually active, and we thus exclude these women in an additional robustness analysis. We added a footnote in the section *Methods - Exclusion Process and Participants* to further explain the decision to exclude them in the robustness analysis:

In PAIRFAM, women who indicated that they had never been sexually active in their life were not asked about their contraceptive method. These women were coded as using no contraceptive method, i.e., a non-hormonal method (see the section about the variables for more information). This coding may introduce some errors as some women may use hormonal methods without being sexually active; we thus exclude them in an additional robustness analysis to ensure that this coding decision does not systematically affect results.

R1.2.a As a follow-up: excluding women who are sterilized, use a nonhormonal IUD, who are trying to become pregnant, or are just not using a listed contraceptive method is excluding a large amount of women who are free from HC use. Comparing the effects of using HCs to only women using certain other methods but not the previously listed circumstances seems a limitation here. Maybe that is the better way to approach this (no study is without limitations) however I would like the authors to consider this potential limitation in the design.

For the reasons outlined by the reviewer, we decided to include women who are sterilized and women who use an intrauterine device in the main analyses, but will still exclude them in additional robustness analyses. Women who are sterilized are classified as non-hormonal contraceptive users in our main analyses (because the most common form of female sterilization, tubal ligation, does not interfere with the menstrual cycle), but their endocrine system might differ substantially from other non-hormonal contraceptive users (e.g., in cases of hysterectomy). The classification of intrauterine device users into the hormonal and non-hormonal contraceptive user group was unfortunately not completely clear (see comment R1.1.).

We believe that the situation concerning women who are trying to become pregnant is qualitatively different. First, we could not rule out the possibility that they were already pregnant without their knowledge. Pregnant women show very different hormonal patterns compared to naturally cycling women, so grouping them with the other non-hormonal contraceptive users leads to bias. Second, there is no reason to use birth control and their sexual activity might be mainly influenced by their wish for children. Third, we are interested in the causal effects of contraceptives, and these are defined by the contrast between different states of the world. We believe that for women who try to conceive, the counterfactual state “trying to conceive, using hormonal contraceptives” is not particularly plausible, thus calling into question whether they can meaningfully contribute to the estimation of a consistent causal effect in our analyses.

For women who indicated that they were using a contraceptive method that was not listed within the options, we were unable to decide whether they were using a hormonal or a non-hormonal method, making it impossible to include them in our

analyses given that in the present study, we are interested in the broad contrast hormonal vs. non-hormonal method (see also response to R1.1.a for our rationale to stick with this general focus).

We hope that these decisions, in combination with extensive robustness checks, strike a reasonable balance between including as many women as possible while ensuring that the analyses still target the substantive effect of interest.

R1.3. I have some concerns about grouping all effects within each DV (sexual desire, sexual behavior, satisfaction; depression, life satisfaction, self-esteem). I would like to see an exploration of each of these individually and one of them comprised into a latent factor/composite of sexuality and well-being. This may be what the authors were intending however this is not specified (and my apologies I am not great at reading other people's code, so I would prefer to see the statistical model included in the report).

Our plan is to only analyze outcomes individually, we apologize that this was not immediately clear from the manuscript. We do not think that it would be feasible to combine the outcomes in the two groups to one latent factor as they are measuring very different aspects of sexuality and well-being. Moreover, we believe that the current project is already very complex, and incorporating additional analyses would only serve to further augment the complexity for the reader to comprehend. Nevertheless, we acknowledge that the research question suggested by the reviewer holds significant promise for future investigations that can build upon the outcomes of the present research endeavor.

To make it easier to follow our planned analysis within the manuscript we added a new section *Models* to the section *Methods*. It includes the equations for the models investigating the potential confounding effects on hormonal contraceptive use, for the adjusted regression analyses, and for the IPTW approach:

Models

All planned analyses can be found in form of an rmd file and an html file:https://osf.io/u8ntf/?view_only=6d5b0a56a41541249cab38c51847157c. In addition, all models are outlined below using simplified readable notation. Code and the notation provided below use the same names for all variables. For the variable outcome, sexuality measures will include desired sexual frequency in the last three months, reported sexual frequency in the last three months, and sexual satisfaction. Well-being measures will include depressiveness, general life satisfaction, and self-esteem. These models are all multilevel models with a random intercept u_{0i} *and in some cases a random slope for hormonal contraceptive use* $u_{1i}hc_{wi}$ *across waves w nested within participants i.* $s(age_{wi})$ *refers to a function of the spline constructed variable for age. Some models are weighted multilevel models indicated by the sign | followed by the name of the respective weight after the outcome variable.* w_miss_i *refers to a weight for systematic missingness reported by PAIRFAM.* w_iptw_i *refers to the weight calculated based on the first step of the IPTW approach.*

Models to Gauge Confounding Effects on Hormonal Contraceptive Use

$$hc_{wi} = b_0 + b_1 * mean\_hc_i + b_2 * hc_{(w-1)i} + b_3 * outcome_{(w-1)t} + b_4 * hc_{(w-1)i} * outcome_{(w-1)t} + b_5 * s(age) + b_6 * log(net\_income_{wi}) + b_7 * years\_edu_{wi} + b_8 * rel\_stat_{wi} + b_9 * rel\_stat_{wi} * rel\_dur_{wi} + b_{10} * rel\_start_{wi} + b_{11} * single\_start_{wi} + b_{12} * n\_kids_{wi} + b_{13} * comp\_fert\_plans_{wi} + u_{0i} + \varepsilon_{wi}$$

with

$$u_{0i} \sim N(0, \sigma_u^2)$$

$$\varepsilon_{wi} \sim N(0, \sigma_\varepsilon^2)$$

Adjusted Regression Analyses

$$outcome_{wi} = b_0 + b_1 * mean\_hc_i + b_2 * hc_{wi} + b_3 * hc_{(w-1)i} + b_4 * outcome_{(w-1)t} + b_5 * hc_{(w-1)i} * outcome_{(w-1)t} + b_6 * s(age) + b_7 * log(net\_income_{wi}) + b_8 * years\_edu_{wi} + b_9 * rel\_stat_{wi} + b_{10} * rel\_stat_{wi} * rel\_dur_{wi} + b_{11} * rel\_start_{wi} + b_{12} * single\_start_{wi} + b_{13} * n\_kids_{wi} + b_{14} * comp\_fert\_plans_{wi} + u_{0i} + u_{1i} hc_{wi} + \varepsilon_{wi}$$

with

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N(0, \Omega_u)$$

$$\Omega_u = \begin{bmatrix} \sigma^2_{u_0} & \\ \sigma_{u_{01}} & \sigma^2_{u_1} \end{bmatrix}$$

$$\varepsilon_{wi} \sim N(0, \sigma_\varepsilon^2)$$

Inverse Probability of Treatment Weighting Approach

Model to Compute Weights w_iptw.

$$hc_{wi} = b_0 + b_1 * mean\_hc_i + b_2 * hc_{(w-1)i} + b_3 * outcome_{(w-1)t} + b_4 * hc_{(w-1)i} * outcome_{(w-1)t} + b_5 * s(age) + b_6 * log(net\_income_{wi}) + b_7 * years\_edu_{wi} + b_8 * rel\_stat_{wi} + b_9 * rel\_stat_{wi} * rel\_dur_{wi} + b_{10} * rel\_start_{wi} + b_{11} * single\_start_{wi} + b_{12} * n\_kids_{wi} + b_{13} * comp\_fert\_plans_{wi} + u_{0i} + \varepsilon_{wi}$$

with

$$u_i \sim N(0, \sigma_u^2)$$

$$\varepsilon_{wi} \sim N(0, \sigma_\varepsilon^2)$$

Outcome Models.

$$outcome_{wi} \mid (w\_miss_i * w\_iptw_i) = b_0 + b_1 * hc_{wi} + u_{0i} + u_{1i} hc_{wi} + \varepsilon_{wi}$$

with

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N(0, \Omega_u)$$

$$\Omega_u = \begin{bmatrix} \sigma^2_{u_0} & \\ \sigma_{u_{01}} & \sigma^2_{u_1} \end{bmatrix}$$

$$\varepsilon_{wi} \sim N(0, \sigma_\varepsilon^2)$$

In addition, we adjusted the code *hc-pairfam_planned-analysis.Rmd* as well.

R1.4. Is there a plan to control for how having never had sex might impact results on sexuality? Perhaps a follow-up analysis with and without these women for sexuality would be appropriate.

We thank the reviewer very much for this comment and agree that this is an important exclusion criterion especially for the sexual frequency outcome as we are conditioning on the outcome (including women that never had sex and are more likely to use a non-hormonal method). We therefore added this as an exclusion criteria to our robustness analyses described in the section *Methods - Exclusion Process and Participants* and in the newly

created Table 2. We expect the subgroup of women who were never sexually active to be too small to benefit from a focused analysis.

R1.5. Figure 1 – while this is an adequate conceptual model, I would like to see more information regarding the statistical model included so that I understand how variables are being treated in analyses. What does relationship mean in this figure? (Relationship status/length?) I would like to see the names of the constructs the same in figures as in tables.

We agree with the reviewer that it was complicated to connect information from the conceptual model outlined in Figure 1 to the variables in Table 3 (former Table 2) and to the statistical models in the code. We do not want to include all variables in the conceptual model shown in Figure 1 as we think this would make the figure too large and hard to follow.

We adjusted Figure 1 slightly to make a clear distinction between conceptual information and variable names (we changed *contraceptive method* to *hormonal contraception* in Figure 1).

In addition, we added a column *Conceptualization in Figure 1* to Table 3 (former Table 2) to connect the variable names to the conceptualization.

To make it easier to follow our planned analysis within the manuscript we added a new section *Models* to the section *Methods* (see our answer to comment R1.3. for more information).

R1.6. Table 1 – please add a column which breaks down how results will be interpreted (especially when multiple modeling approaches will be used to answer the same questions) per the PCI RR recommendations.

We thank the reviewer for this comment and added a column “Interpretation given different outcomes” to Table 1. For the analyses identifying the average treatment effects of hormonal contraceptive use on sexuality and well-being we added the following rule to this column:

If outcomes based on the two estimations differ, adjusted linear regression analyses will be treated as the main analysis and the inverse probability of treatment weighting approach will be treated as a robustness analysis for identifying the average treatment effect

For all other estimands we will use only one modeling approach. As we are focusing on estimands and are not interested in (dis)confirming hypotheses we have not included the column “Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis”.

R1.7. For the heterogeneity question – I would love to see this question addressed with more statistical nuance than comparing the percentage of women with positive effects (guessing effects here is meant to refer to the totality of sexuality effects and well-being effects separately, but the authors do not adequately explain how they plan to conceptualize these comparison variables) to the percentage with negative effects. Here, again, I would be interested to see if there is heterogeneity in responses to treatment in each DV and the latent/composite DVs. I would also be curious if there was a way to profile those with different effects (not just positive or negative) based on the groupings of their responses to treatment. (e.g., is there a group of women with high self-esteem and sex drive, but low sexual behavior? Another group with high sexuality and high depression? Is there something that predicts membership to these groups?)

As described in our answers to comments R1.3 and R2.6, we agree with the reviewers that combining the outcome measure would be a very interesting research question in itself. Nevertheless, the current manuscript is already very complicated and aims to answer several complex questions at once. Therefore, we would like to keep the additional analyses concerning individual treatment effects as straightforward as possible. Nevertheless, we acknowledge that the research question suggested by the reviewer holds significant promise for future investigations that can build upon the outcomes of the present research endeavor. Even though we particularly agree with the reviewer that it would be interesting to investigate the relationship between individual treatment effects across the two outcome groups, we think that this is not feasible as the programmatic registered report will be published in two stage 2 articles (one stage 2 article focussing on sexuality and one stage 2 article focussing on well-being).

R1.8. For the question of investigating if adverse experiences on HCs influence future use – how does investigating relationships between how long a woman has been using HCs and her future use answer this question? Are there additional variables being included in the model to answer this question that I am missing? How are positive or negative experiences being conceptualized?

We think that the reviewer is referring to our analyzes investigating the link between individual treatment effects and contraceptive decisions. Reviewer #2 also asked for clarification concerning this research question in comment R2.13.

The research question is whether women use their own experience with individual effects of hormonal contraceptives on sexuality and well-being to make a decision about their contraceptive method (e.g., women who experience adverse effects of hormonal contraceptives on sexuality or well-being might be more likely to stop using them; this is described in the section *Heterogeneity in Treatment Response* in the introduction). As further discussed in the section *Estimands* in the introduction, we want to correlate *years using hormonal contraceptives* with the individual treatment effect (i.e., positive or negative experiences with hormonal contraceptives).

To make it more explicit that we are not interested in predicting future behavior but rather see this analysis as a first indicator for potential assortment based on experiences we revised the following sections.

Heterogeneity in Treatment Response in the introduction*: For example, are women who experience adverse effects of hormonal contraceptives on sexuality or well-being more likely to stop using them during a specific time span?*

Estimands in the introduction*: Ideally, we would have sufficient data to instead estimate individual treatment effects (e.g., using all but the last wave of data) to predict individual behavior (e.g., contraceptive method in the very last wave of data). However, in the context of the available data, this would result in very low statistical power, and we thus decided on a different approach which would only provide very rough evidence for potential assortment based on experiences with contraceptive methods. Such an assortment based on experiences would result in the type of selective attrition explained above and may provide a partial explanation for the mixed evidence concerning effects of hormonal contraceptives on sexuality and well-being.*

Methods - Analysis Plan: *To answer the question whether women guide their contraceptive method choices by deciding against hormonal contraceptive methods after experiencing adverse effects, we will again use individual treatment effect estimates from the adjusted regression analysis, this time correlating them with the proportion of years using hormonal contraceptives (waves in which hormonal contraceptives were used divided by total number of waves participating in PAIRFAM). This correlation analysis will again be weighted by the inverse of the standard error of the individual treatment effect estimates. This analysis can potentially provide tentative evidence for assortment based on experiences with contraceptive methods.*

We hope that this additional information answers the questions by reviewer #1 and reviewer #2 and helps the reader to follow the idea of our analyses investigating contraceptive method choices.

R1.9. For data simulations, how was -.45 selected as the estimate of the true causal effect?

The size of the underlying causal effect was chosen arbitrarily and for simulation purposes only. Varying it would add a lot of additional complexity to the simulation. Nevertheless, it is not completely out of line: A recent double-blind, randomized, placebo-controlled trial found effect sizes of oral contraceptives on sexuality ranging from *Cohen's d* = -0.24 to -0.22 (Zethraeus et al., 2016) and on well-being ranging from *Cohen's d* = -0.41 to -0.22 (Zethraeus et al., 2017).

We added this information to the section *Methods - Simulation*: *The size of the underlying causal effect was chosen arbitrarily for the purpose of this simulation. Nevertheless, it is not completely implausible; a recent double-blind, randomized, placebo-controlled trial found effect sizes ranging from Cohen's d = -0.41 to -0.22 (Zethraeus et al., 2016, 2017).*

R1.10. Why were the following variables not considered here: age at first HC use, duration of HC use prior to the study beginning, childhood trauma, pre-existing mental health Dx. If data not available, this is absolutely appropriate, however, given that these factors are more supported as influencing HC treatment outcomes than are personality traits, I would expect these to be included.

We do agree that it would be very interesting to correlate some of the proposed variables with the individual treatment effects and thank the reviewer for the suggestions.

Unfortunately, information about age at the first hormonal contraceptive use and duration of hormonal contraceptive use before the start of the study is not available in PAIRFAM. (This also means that in other analyses we can only rely on a proxy for duration of hormonal contraceptive use. For example, we therefore decided to correlate the proportion of years using hormonal contraceptives during the course of PAIRFAM divided by the total years participating in PAIRFAM with the individual treatment effects to investigate whether women's decision which contraceptive method to use is informed by their individual treatment effects (i.e., women who experience very negative individual treatment effects when using a hormonal contraceptive method are more likely to use a non-hormonal contraceptive method during the course of PAIRFAM).)

While there is some evidence that childhood trauma is associated with an increased risk of hormonally mediated syndromes such as PMDS and PMDD (for a short summary see Hill & Mengelkoch, 2023), we found no research linking childhood trauma to the effects of hormonal contraceptives on sexuality or well-being (even though we agree that they might influence the outcomes itself but this is not relevant for the question how interindividual differences are connected to individual treatment effects that we aim to answer here). Hill & Mengelkoch (2023) argue that the effect of hormonal contraceptive use on women's mood may be moderated by childhood trauma in a form that women with early stress exposures have an increased vulnerability to negative side-effects, particularly those related to mood. However, PAIRFAM only collects information about participant's general satisfaction with their childhood and specific traumas during later life (and the latter only in wave 7, so for a much reduced subsample). We agree with the reviewer that this would be an interesting question for future research but decided we could not address it with the PAIRFAM data.

For pre-existing mental health, there is information about mental illness and addiction in PAIRFAM but this information is also only available from wave 7 upwards. Again, we do not have a strong hypothesis on how mental health might be connected to individual treatment effects. In addition, we also do not see a convincing reason to include them as confounders, which would lead to a lot of missing data.

R1.11. Given the multilevel nested nature of this data (and my own training), I might have approached analyses using a RI-CLPM. I would love if the authors could describe why they selected their analysis plan in comparison to this type of analysis plan (or other appropriate models)! That is, the authors explained why their approach was a good approach, but not why it might be a better way to answer this question compared to other approaches! I am very ready to be convinced.

Thank you for the excellent question. The bivariate RI-CLPM does indeed control for time-invariant confounders in a manner similar to the fixed effects approach we employ. But there are several reasons why we settled on the latter model.

First, we additionally want to take into account time-varying confounders, and incorporating these into a RI-CLPM greatly increases model complexity (it is definitely possible, but as far as we can tell, it usually isn't done).

Second, and substantively more importantly, in a RI-CLPM the focus is usually on the lagged coefficients. Given the large gaps between assessments in our data, we do not believe that these provide the best fit to the substantive effect of interest, as we are more interested in the immediate effects of hormonal contraception (rather than the effects of last year's contraception). Of course, hormonal contraception may have long-term effects (in particular if taken over longer periods of time), but identifying them would call for a different substantive estimand and a different data-analytic approach (not necessarily a RI-CLPM either).

Third, the RI-CLPM (at least in the out-of-the-box version) focuses on bidirectional associations; but here, we have little direct interest in, for example, how well-being affects choice of contraceptive. The reverse questions are of course interesting in their own right, but not the focus of the present manuscript. Given our interest in concurrent effects of contraceptive use on the included outcomes, the fixed effects multilevel-regression model is the most "lightweight" approach that does not add any superfluous complexity. We believe that focusing on a narrower set of questions increases the chances that we can provide a valid and thorough answer.

As you have pointed out, this choice (which is aligned with practices in fields more focused on causal identification) may be surprising to readers trained with more focus on SEM. We have thus added the following justification to the section *Conceptual Design and Underlying Assumptions* in the introduction:

Given our interest in the immediate effects of hormonal contraceptive use (rather than the lagged effects after one year), and to avoid adding superfluous complexity, we decided against a popular alternative modeling approach (RI-CLPM, Hamaker et al., 2015) which simultaneously attempts to estimate causal effects pointing into the opposite direction.

Reviewer #2's remarks:

This manuscript presents the Stage 1 proposal to examine relating to effects of hormonal contraceptive use and two outcome measures (sexual desire and well-being, which will be presented in 2 separate Stage 2 manuscripts). The research question is important and I commend the authors for finding a dataset that can be used to test their hypotheses (though I understand that the research is not actually testing hypotheses as much as describing the strength of associations). This will be an important contribution to the literature. I did find some aspects of the Stage 1 proposal unclear and/or difficult to follow. I list specific concerns below, ordered by section.

We would like to thank the reviewer for their helpful and constructive feedback. Based on the specific concerns we clarified our theoretical considerations (see responses to R2.1., R2.2, and R2.3), the included variables (see responses to R2.9., R2.10, and R2.11), and the statistical model (see responses to R2.4., R2.5., R2.12., and R2.13.). We now provide more information on the exclusion process (see responses to R2.7., R2.8., and R2.9). We have addressed all the specific concerns in more detail below.

Introduction

R2.1. I found the logic transition from talking about the need to conduct experiments to examining treatment effects (in the first part of the introduction) to be muddy. I think the authors can create a stronger framing for their use of the longitudinal data. I would de-emphasize experimental approaches and just focus on the strength of the present approach, which is strong in its own right. I am also then unsure about the focus on randomized control trials right after listing the five possible explanations for the mixed findings in the literature. Again, I would reframe to focus on the current method. If a comparison to randomized controlled studies is needed, I would add it after.

We thank the reviewer for this comment. As experiments are considered the gold standard to answer causal research questions (some argue they are the only option) we wanted to stress the importance of other research designs for causal inference early in the manuscript. To make it easier for readers to follow our reasoning, we added the following explanation early in the introduction:

Experiments are considered the gold standard to answer causal research questions such as the effects of hormonal contraceptives on sexuality and well-being. However, experimental evidence can only partly tell us how these effects affect women's everyday lives.

Considering the five possible explanations for the mixed findings in the literature, we think that it is important to clearly discuss which explanations are addressed by randomized controlled trials and which aren't. Therefore, we would like to keep the structure here as it is to first discuss randomized controlled trials and their limitations. The next section labeled *Observational Cross-Sectional and Longitudinal Designs* in the *Introduction* then focuses on the current method extensively.

R2.2. I had to read this sentence over multiple times before I understood what exactly was being said because it seems to start a completely different point than the beginning of the paragraph (the sentence that starts with "To apply this perspective to the current research question, assume that for example an ineffable..." on page 11). I think the authors can make this point in a more straight-forward way. Perhaps by introducing the time-varying versus time-invarying terms first.

We thank the reviewer very much for this feedback and have restructured this paragraph following their suggestion to introduce time-invariant and time-varying confounders first:

Given the correct modeling decisions, time-invariant confounders are automatically controlled for in longitudinal designs. As they do not vary within a woman, they will not induce spurious correlations between her time-varying predictor and her time-varying outcome. Time-varying confounders on the other hand are not automatically controlled by longitudinal designs, but instead need to be accounted for (Rohrer & Murayama, 2021). A time-varying confounder might affect a woman's choice of contraceptive method as well as the outcome of interest at a given time. For example, an ineffable or at least unmeasured shift from a casual to a more steady exclusive relationship may affect the decision to use hormonal contraceptives. In addition, this shift could cause more frequent sexual activity at a later time. In a longitudinal design that only measures hormonal contraceptive use and sexual activity but not this relationship shift, it will appear like there is a positive causal effect of hormonal contraceptives on sexual activity.

R2.3. Skovlund et al., (2016) found age to be particularly important for whether HC use was associated with greater risk for depression. I think this point should be more thoroughly discussed. Additionally, are any of the findings in the sexuality section similarly moderated by age? The Skovlund reference may also need to be mentioned after the "Older women might be more likely to experience beneficial side effects.." on page 14. The given reason for this association focuses on finding a method that fits them best, however, is there any evidence that sensitivity to steroid hormones decreases with age as well?

Even though we do not know of any evidence that sensitivity to steroid hormones decreases with age, we do agree with the reviewer and Skovlund et al. (2016) that age might be an interesting moderator for the effect of hormonal contraceptive use on sexuality and well-being.

This is one of the reasons why we decided to include age in our analyses investigating the correlations between interindividual differences and individual treatment effects. Estimating the correlation between individual treatment effects and age is conceptually the same as including age as a moderating variable in the analyses. We do think that estimating correlations between interindividual differences and individual treatment effects will allow us further insights into the heterogeneity in treatment responses.

We have added some more information around the Skovlund et al. (2016) reference to the *Heterogeneity in Treatment Responses* in the introduction to connect their work with our idea of investigating individual treatment effects and interindividual differences.

In addition, we want to answer the question whether interindividual differences like demography and personality predict individual treatment effects. Older women might be more likely to experience beneficial side effects of hormonal contraceptives on sexuality and well-being because they found the method that fits them best. In line with this reasoning, empirical findings suggest that higher age was associated with less negative side effects of hormonal contraceptive use on depression with particularly strong negative effects during adolescents (Skovlund et al., 2016). Nevertheless, these findings might be accountable by other explanations, e.g. a possible decrease in sensitivity to steroid hormones with age or a specifically strong sensitivity to steroid hormones during puberty.

We are not aware of any evidence showing a similar moderating effect of age on the association between hormonal contraceptive use and sexuality. We thank the reviewer very much for this great suggestion.

<u>Conceptual Design and Underlying Assumptions</u>

R2.4. In both the time-lagged regression and the IPTW approach, the model includes the interaction of the outcome and contraceptive method at the previous assessment (in addition to the main effects). It is not clear to me conceptually why the interaction term is needed. Can the authors clarify their hypothesis? If the interaction term is not significant, will it remain in the model?

We included the interaction term in the IPTW approach to model the possibility that certain outcomes might have stronger effects in hormonal contraceptive users than in non-hormonal contraceptive users on the contraceptive choice. For example, it is very likely that strong negative side effects on sexuality and well-being will lead to the decision to stop using hormonal contraceptives as they might be attributed to the contraceptive method. We include this interaction term in the adjusted regression analyses to keep them parallel. We will not perform stepwise variable selection and all predictors will remain in the model (even if they are non-significant).

To explain the decision to include this interaction in our models, we added a footnote in the section *Methods - Analysis Plan*:

We decided to include the interaction term in the IPTW approach to model the possibility that certain outcomes might have stronger effects in hormonal contraceptive users than in non-hormonal contraceptive users on the contraceptive choice (e.g., strong negative side effects on sexuality and well-being might be more likely to be attributed to the contraceptive choice in hormonal contraceptive users leading to the decision to stop using this method). To keep both approaches parallel, we also included this interaction term in the adjusted regression analyses.

R2.5. For variables related to relationship status, would it not be best to include relationship status at previous assessment to examine change in relationship status? This seems like a better test of the theoretical rationale that was laid out (about starting a relationship being a potential confounder).

We thank the reviewer for this idea and agree that this is an important confounder. We now include two dummy-coded variables in all analyses, one indicating whether a woman started a relationship between the previous wave and the current wave and one indicating whether a woman became single between the previous wave and the current wave to control for the potential confounding effect.

To incorporate this into the manuscript, we added a row to Table 3 (former Table 2) in the section *Methods - Variables* to include the current relationship status. In addition, we added the following sentence to the section *Methods - Analysis Plan*: *In addition, we will include two dummy coded variables: one indicating whether a woman started a relationship between the previous wave and the current wave and one indicating whether a woman became single between the previous wave and the current wave.* To make it easier to follow our planned analysis within the manuscript we added a new section *Models* to the section *Methods* (see our answer to comment R1.3. for more information). Finally, we adjusted the code *hc-pairfam_planned-analysis.Rmd* as well.

R2.6. Would the data allow for some sort of latent class analysis to classify women as having positive, negative, or neutral effects?

We thank the reviewer very much for this idea. Reviewer #1 proposed something similar in their comment R1.3. As outlined in our answer to comment R1.3. and R1.7., we think that adding more complex analysis to the current research project would make it even harder for the reader to follow the analyses focussing on separate outcomes and heterogeneity in these effects.

<u>Method and Analysis Plan</u>

R2.7. Do the authors mean that women will be excluded if they ever hit 50 or menopause during the data collection period or only their timepoints after 50/menopause (i.e., earlier timepoints from those women would be included)? That sentence was unclear to me.

The lack of a clear explanation of our exclusion criteria for our main and robustness analyses and specific reasons for these criteria were also criticized in comments R1.1., R1.2., R1.2.a, R2.8, and R2.9. Therefore, we rewrote the section *Methods - Exclusion Process and Participants* and added Table 2 summarizing exclusion criteria, reasons for exclusion, and unit(s) that will be excluded.

We apologize that the sentence about exclusion criteria for current and subsequent waves was unclear. Women, who did not identify as female, were older than 50, or (post-)menopausal will be excluded for the current and all subsequent waves but **not** for previous waves. To make this clear we added that previous waves will be included in the

analyses in the section *Methods - Exclusion Process and Participants*: *We plan to exclude individuals who did not identify as female. Furthermore, once a woman crossed the age of 50 or reported to be (post-)menopausal, her data (including subsequent waves) will be excluded, but previous waves of data collection will remain in the analysis.*

R2.8. What is the logic for excluding women in homosexual relationships? Research on cycle effects documents similar shifts in women who are in homosexual relationships as women who are in heterosexual relationships.

The exclusion of homosexual relationships was also criticized by reviewer #1 in comment R1.2.

We are not aware of any scientific evidence comparing cycle effects in heterosexual and non-heterosexual women except for a study by Diamond & Wallen (2010) based on a very small sample size (N = 20). From a theoretical perspective we find it plausible that non-heterosexual women experience similar cycle shifts as heterosexual women. Nevertheless, the reasons for using contraceptive methods differ for homosexual women as they do not have to avoid unwanted pregnancies (and will probably use hormonal contraceptive methods less often). We think that the decision process which contraceptive method to use will therefore differ notably for homosexual women. This is the reason why we decided to exclude waves in which women were currently in a homosexual relationship or exclusively reported homosexual relationships in the past. Unfortunately, sample sizes for women in non-heterosexual relationships will be too small to reliably reach any conclusion given the expected small effect sizes.

R2.9. I am not sure why women who indicated using no contraceptive method will be excluded in robustness analyses. That sentence does not make sense to me. Specifically, after reading this sentence, I became confused about the "control group" - will it just be users of NON-HC METHODS or anyone not currently on a HC? NON-OC METHODS? The authors should specify exactly how hormonal contreceptive use will be operationalized from the questions in the PARFAIM dataset (if this is reported, it is not prominent enough).

As reviewer #1 also notes in their comments R1 and R1.1.a that it was hard to follow how our variable hormonal contraception was built, we describe the transformation of this variable in more detail in the section *Methods - Variables*:

The predictor hormonal contraception will be based on the items about the contraceptive method; participants were able to report multiple contraceptive methods. Hormonal contraception will be coded as 0 if participants indicated that they use no contraceptive method at all. The variable hormonal contraception will also be coded as 0 if participants indicated that they use no hormonal contraceptive method and at least one of the following methods: condom*;* intrauterine device[4]*;* diaphragm, foam, suppository, gel*;* natural birth control*;* female sterilization*;* male sterilization*; or* withdrawal method, coitus interruptus*. In addition, the variable hormonal contraception will be coded as 0 if participants were never sexually active in their life, as these participants were not asked about their contraceptive*

method. The variable hormonal contraception will be coded as 1 if participants indicated that they use a birth control pill*,* mini pill *or* other hormonal method (implant, patch, ring) *even if they additionally use non-hormonal methods. Exclusion criteria for main as well as robustness analyses based on the contraceptive method are described above.*

To answer the question raised above, this means that the control group will include women that use non-hormonal contraceptive methods as well as women that use no contraceptive method. We decided to exclude the group of women using no contraceptive method from the robustness analysis to check whether obtained results based on the main analyses were only driven by the subgroup of no contraceptive method users that presumably differ in further aspects from other non-hormonal contraceptive users.

R2.10. Are there questions on whether women had sex in the last 3 months? Or whether women are sexually active?

Yes, one of the outcomes (reported sexual frequency) refers to the sexual activity in the last three months. The exact wording can be found in Table 3 (former Table 2). PAIRFAM also includes the information whether women had been sexually active for the first time (see comment R1.4 for a detailed description on how we plan to include this information in our study). There are no further variables measuring sexual frequency available in PAIRFAM.

R2.11. Is there any information on age of menarche?

Unfortunately, this information is not available in the PAIRFAM dataset.

R2.12. Why is relationship duration separated into quantiles?

We originally separated relationship duration into quantiles to include singles in our analyses in a straightforward manner. Assigning singles a relationship duration of 0 months seemed unfeasible, given that they potentially differ in a non-linear form from participants currently in a relationship; turning the variable into a categorical one seemed like an easy solution.

But in the meantime, we have learned of a different approach for these nested variables that are missing for a specific subgroup to be included in an analysis. This information can be included by entering a dummy coded variable for the missingness (in our case being in a relationship or being single) and adding an interaction of this dummy-coded variable with the variable of interest (in our case log transformed relationship duration). The variable of interest is set to an impossible value for the subgroup with missing information (in our case we will set relationship duration to -1 for all singles). The important part of this approach is that the main effect of the variable of interest (i.e., relationship duration) is not included in the analyses.

We thus decided to include relationship duration as a nested variable in our analyses. We explain this in detail in the section *Methods - Analysis Plan*:

Furthermore, relationship duration will be included as a nested variable. This allows us to model a linear association with relationship duration which is only informed by women who are in a relationship, while simultaneously including those who are not in the analysis. Technically, we achieve this by including a dummy coded variable for current relationship status (single vs. non-single) and its interaction with log transformed relationship duration as a predictor. No main effect of relationship duration will be included in the model. Relationship duration for singles will be set to -1; this value is arbitrary and does not affect the resulting estimates because when multiplied with the relationship status dummy, relationship duration for singles is dropped from the analysis. In addition, we will include two dummy coded variables: one indicating whether a woman started a relationship between the previous wave and the current wave and one indicating whether a woman became single between the previous wave and the current wave.

To make it easier to follow our planned analysis within the manuscript we added a new section *Models* to the section *Methods* (see our answer to comment R1.3. for more information). Finally, we adjusted the code *hc-pairfam_planned-analysis.Rmd* as well.

R2.13. For the test of whether women guide their contraceptive method choices, will the authors use HC use at the last wave as a dichotomous DV? I found that paragraph a bit difficult to follow.

We think that the reviewer is referring to our analyzes investigating the link between individual treatment effects and contraceptive decisions. Reviewer #1 also asked for clarification concerning this research question in comment R1.8.

To make the paragrapher easier to follow (we think that the reviewer is referring to the paragraph describing our analyses for investigating contraceptive method choices in the section *Methods - Analysis Plan*), we rephrased the section and provided more information:

To answer the question whether women guide their contraceptive method choices by deciding against hormonal contraceptive methods after experiencing adverse effects, we will again use individual treatment effect estimates from the adjusted regression analysis, this time correlating them with the proportion of years using hormonal contraceptives (waves in which hormonal contraceptives were used divided by total number of waves participating in PAIRFAM). This correlation analysis will again be weighted by the inverse of the standard error of the individual treatment effect estimates. This analysis can potentially provide tentative evidence for assortment based on experiences with contraceptive methods.

In addition, we added information to the following sections to further explain the idea behind our analyses investigating contraceptive method choices:

Heterogeneity in Treatment Response in the introduction*: For example, are women who experience adverse effects of hormonal contraceptives on sexuality or well-being more likely to stop using them during a specific time span?*

Estimands in the introduction*: Ideally, we would have sufficient data to instead estimate individual treatment effects (e.g., using all but the last wave of data) to predict individual behavior (e.g., contraceptive method in the very last wave of data). However, in the context*

of the available data, this would result in very low statistical power, and we thus decided on a different approach which would only provide very rough evidence for potential assortment based on experiences with contraceptive methods. Such an assortment based on experiences would result in the type of selective attrition explained above and may provide a partial explanation for the mixed evidence concerning effects of hormonal contraceptives on sexuality and well-being.