Reviewer 3 Revision review

I read the response to reviewers and re-read the manuscript. I was satisfied with the authors responses to my queries and Dr. Rohrer's queries. I was surprised by the novel section on the historical roots of the BO research. I would make this a separate paper, as it has no strong connection with the rest of the manuscript, which is already very long. It's certainly interesting, but who knows, maybe the controversial/"political" aspect will overshadow the methodological aspect and provide something for detractors to focus on, even though the case against OBOR is clear no matter where you stand on citing Nazi science. My recommendation is that the authors separate this out instead of "tacking it on", but it's their manuscript. It's also worth noting that Ray Blanchard has written a rambling reply to the preprint of the manuscript under review here. https://link.springer.com/article/10.1007/s10508-022-02362-z I think he confused several key issues, perhaps in part because the manuscript was very long and a bit hard to read, perhaps in part because he is very invested in his theory. I don't think the authors need to reply to Blanchard 2022 here, as it needlessly makes the debate even more confusing to explain the confusions related to a previous version before rebutting his other points. I am quite hopeful that the revised version is a slightly easier read.

I hope that we will soon see better research on the important topic of the causes of homosexuality thanks to this manuscript. I made some minor responses within the text.

Response to Reviewers

Dear Editor, Dear Reviewers –

Thank you for your detailed and helpful remarks on the manuscript, which, given its length, must have been a lot of work. The reviewers pointed out important issues which we were entirely unaware of, and it took us some time to figure out a way of how to best incorporate these new insights into the manuscript. Most importantly, the previous versions of the manuscript were received as if we wanted to make substantive or causal claims about the FBOE (this became clear to us from the remarks of all three reviewers). Yet, we never intended to make any causal or substantive claims about the origins of sexual orientation, or the biological explanation put forth by proponents of the FBOE. Our perspective is this: In their research, Blanchard et al. put forth a set of assumptions (these include the MIH and the claim that each older brother causes the odds of homosexual orientation to increase) from which they derive the necessary observation that homosexual men should have more older brothers, but also that they should have slightly more older sisters. Importantly, this group difference in the number of older sisters should be smaller than the group difference with respect to older brothers, and in most of the literature it is claimed that any difference in the number of older sisters arises due to older brothers and older sisters being positively correlated and that this difference does not have any substantive meaning on its own. Regardless of what factors cause these differences in the number of older brothers and sisters to occur, the relevant observation which the FBOE rests upon (i.e., the theoretical estimand) is the difference of the difference in the number of older brothers and the difference in the number of older sisters (i.e., the surplus of older brothers relative to older sisters in homosexual vs. heterosexual men). Throughout the paper, we were primarily concerned with the methods currently employed for detecting/assessing this "difference of differences". Our aim was to show that these methods and practices are incapable of detecting this difference of difference and thus lead to artefactual, erroneous conclusions. We went over the manuscript multiple times, changing some of the words we used to better distinguish that we are first and foremost not concerned with causal claims (e.g., we replaced all instances of "controlling for…" by "adjusting for…")

We hope that our responses below and the changes we made in ms. revision adequately address the issues you raised. Among many other amendments (please refer to our point-by-point responses to the throughout very helpful and constructive comments and suggestions provided by the reviewers), we now refer to the very recent (July 2022 journal print issue), extremely large (population-based) and thus important FBOE research paper by Ablaza et al. (2022). The empirical evidence of this new report fully supports our conclusions derived from our method-triangulating approach to the FBOE literature (namely, convergent evidence based on algebraic derivations utilizing probability calculus, on a data-simulation study, and on a multiverse meta-analysis of the pertinent literature). Further, in the revised ms. we have appended a Coda to the Discussion section, therein highlighting widely unknown and unacknowledged roots of FBOE research in National Socialist pseudoscience (from the mid-1930s to early 1940s) on "hereditary homosexuality", which forerunners are relevant and informative for the current state as well as the future of inquiry along these lines.

Reviewer 1: Julia Rohrer

Basic reporting

Most of the reporting is provided in clear, unambiguous and professional English. I think there is some room to optimizing phrasing in some parts, more on that in my general comments.

Experimental design

The research question/central arguments seem well defined, relevant and meaningful. The research clearly fills/addresses a knowledge gap/issue.

Validity of the findings

In general, I do not see any issue with the validity of the findings reported. I will refer to the conclusions in more detail in my general comments.

Additional comments

This article presents simulations to argue that the so-called fraternal birth order effect (older brothers increase the odds that a man reports a homosexual orientation) may be a statistical artifact. I will partially structure my review according to the points suggested by PeerJ and try to address all checks I'm supposed to provide. However, you will find the bulk of my comments under "General Comments" – sorry for that, this is just how I structure my reviews (and also how I like to receive reviews as an author). Also, I do not understand why I would paste my review into fields that are not sent to the authors (I'm writing the review explicitly with the authors in mind).

Generally speaking, I believe that this manuscript has great potential to make a meaningful contribution to the literature, clarifying a topic that has attracted some attention and lead to a considerable degree of confusion. I reviewed one of the discussed meta-analyses by Blanchard, so I got a brief glimpse of the literature, and it was really frustrating that nobody had fully addressed the underlying methodological issues. So, from my perspective, this manuscript is urgently needed to fill a gap in the literature; and I commend the authors for working their way through this literature (which seems to be a bit of a mess, in particular the more technical claims by Blanchard himself). My review focuses on ways that the manuscript could be improved to ensure that it is read and understood by the broadest possible audience. Also, I will only talk about Part 1 and Part 2. That is mainly because I'm running out of time and will be on vacation for the next two weeks. Sorry for that! I believe that my comments already impose quite a bit of work on the authors, and I'm happy to take a look at Part 3 during the revision (or maybe another reviewer will review that part in more detail).

Best regards, Julia Rohrer (I sign all my reviews)

Response: Thank you for all these positive remarks!

General Comments Major:

Starting from line 200: I can't quite see yet how you move from (1) to (2). I understand that the numerator must be greater than the denominator, but these do not only contain P(PB|…) but also the other part (1 – P(OB|…)). Maybe there is an obvious reason why you could drop this, but it wasn't obvious to me. Could you please explain?

Response: Thanks for pointing this out to us. We see that the justification of (2) is missing. There may be several ways to obtain (2) from (1). To us, the most straightforward way is algebraically:

$$
\begin{aligned}
\frac{\dfrac{P(OB|Hom)}{1-P(OB|Hom)}}{\dfrac{P(OB|Het)}{1-P(OB|Het)}} &> 1 \\
\Leftrightarrow \frac{P(OB|Hom)}{1-P(OB|Hom)} &> \frac{P(OB|Het)}{1-P(OB|Het)} \\
\Leftrightarrow P(OB|Hom)[1-P(OB|Het)] &> P(OB|Het)[1-P(OB|Hom)] \\
\Leftrightarrow P(OB|Hom)-P(OB|Hom)P(OB|Het) &> P(OB|Het)-P(OB|Hom)P(OB|Het)
\end{aligned}
$$

Now both sides of the inequality contain the term $-P(OB|Hom)P(OB|Het)$. Thus, by adding $+P(OB|Hom)P(OB|Het)$ to both sides, we have the desired result

$$P(OB|Hom) > P(OB|Het).$$

We added a more readable algebraic derivation of Equation 2 to the online supplement.

Application of the law of total probability: Is…is that really the quickest and most intuitive way to get to equation 3? Maybe it is. I suspect you might lose readers here though (at least in my experience, psychologists tend to struggle with…really any sort of math). I guess if I needed to explain this, I may initially drop the conditioning on sexual orientation (and think about siblings of homosexuals and siblings of heterosexuals as separate samples). Then it's really just the definition of a conditional probability re-arranged: P(OB|Older) = P(OB, Older)/P(Older) = P(OB)/P(Older). Not sure whether that makes sense to you. In any case, you may keep the derivation the way you did it, but I think you should either add some more verbalization to explain it, or simply provide a worked example with some numbers.

Response: To us, the problems inherent to the OBOR (and similar metrics) became only apparent, once we had abstracted it using the familiar notation and results from probability calculus. The law of total probability affords a level of clarity and rigor which we believe to be unobtainable through intuition alone and so far has been lacking in the FBOE research literature.

The explanation you give makes a lot of sense to us, and it is precisely what we intended to express using Equations 3 and 4.

We could drop the condition of sexual orientation in our notation, but this would mean that we would have to find another way of distinguishing between the OB's in the two groups, say using subscripts $P(OB_{\text{Hom}}|Older)$ and $P(OB_{\text{Hom}}|Older)$, or $P_{\text{Het}}(OB|Older)$ and $P_{\text{Hom}}(OB|Older)$. These would be just alternative ways of expressing the conditional information of which group the respective siblings belong to.

However, we do think that verbalizing these equations is a great idea. Accordingly, we have amended Part I to contain verbalizations of Equations 1-4. We hope that this helps readers to reason about these equations, even if they are unfamiliar with the notation.

Starting from line 234: I find this illustration with numbers nice. However, as somebody coming from the substantive side, what I mainly read into what you are saying is: if we assume some birth order effect on sexual orientation (homosexual men having more older siblings), than we get a positive OBOR even if there is no excess of older brothers relative to older sisters. I think it might be helpful to spell this out? A birth order effect on sexual orientation may still be interesting (maybe). In general, I see that you are coming from the technical side, but I think for readers it is more helpful to have some guidance on what these numbers imply about reality.

Response: Yes, this is what we intended to communicate. If homosexual men tend to have more older brothers *and* sisters, then the OBOR cannot be used to claim that homosexual men have *just* more older brothers as is done in the literature. This was not clearly enough stated by us. We now elaborate on this thought experiment by clearly stating the respective assumptions.

Equation 6: Are you fully being serious about using the OR as a metric? I'd only suggest it to people who really, really, love Odds Ratios. I'd probably just compare the conditional probabilities P(OB|Older) between the two groups (the probability that an older sibling is a brother in homosexual men vs heterosexual men). Or, even better: estimate the effect of older brother (vs. older sister) on the probability of homosexuality, because that is probably how people think about the effect in their heads.

Response: We agree that the odds ratio is not as intuitive as other metrics (and we provide a meta-analysis using the difference in proportions as a metric in the supplementary materials). We think that we did not articulate this with enough clarity in the originally submitted ms., but the odds ratio in Equation 6 fixes the OBOR, assuming that researchers love odds ratios so much that they cannot resist using them. We added a sentence stating that the difference in proportions could be used as well. This antedates the meta-analysis in Part III, but given the available data (or lack thereof), there is only a limited number of options for adequate effect size metrics. We fear that the OR, or, equivalently the difference in proportions are the only "user-friendly" alternatives (in the sense that this way a recipe-like textbook meta-analysis can be carried out on the data) here.

We wholeheartedly agree that it would be better to estimate the effects of older brothers/sisters on homosexuality, but that would require individual-level data, and such data are not available. We think that we needed to state more clearly (and hopefully succeeded) that our reasoning is that if one is interested in whether homosexual men have more older brothers (relative to older sisters) than heterosexual men, then the OR would be a better way to do it.

Line 293: I do not understand what you mean by a shift from the level of participants to the siblings reported by these participants. Wasn't the data hierarchically organized before as well? Why would the OR but not the OBOR violate some assumption? I have a suspicion of what you are trying to get at, but it is really not clear to me. If you wanted to do an analysis on the level of the participant (which I would greatly prefer), it would be something like: Homosexuality ~ Number of older brothers + number of older siblings (that is, you compare people with the same number of older siblings but varying fractions of older brothers). (in this analysis, the effect of older brothers would not necessarily be causally identified due to endogeneous selection bias, but that is a bit of a separate question) Note on the MROB: Did Blanchard really sort of try to reinvent statistical control? Oh boy, this literature is wild.

Response: This is an issue that - thankfully - all three reviewers pointed out to us, and we had to reconsider the argument we made. The primary data used throughout this literature comprises mostly homosexual and heterosexual individuals (but also other categories of sexual orientations can be found here and there) who reported the number of older brothers, older sisters, younger brothers, and younger sisters. An analysis at the level of the individuals whose data was collected would be - as in your example - to model sexual orientation as an outcome variable and number of older brothers/sisters/whatever as covariates/predictors. (But such data are not available)

In using the OBOR or the OR in Equation 6, the individuals who were asked how many brothers and sisters they had are no longer the units of analysis, but instead the following is implied by its use (and this is more or less forced upon anyone trying to use the available data):

Say, a homosexual individual in a given sample reports to have 2 older brothers, 2 older sisters, 1 younger brother, and 1 younger sister. Via the OBOR (and also the OR) Blanchard et al. now would treat each of these siblings as independent observations of the binary variable "older brother" (OB), which is equal to 1, if the reported sibling is an older brother, and equal to 0 otherwise. So in this example the sample size changes from 1 (the individual) to 6 (the individual's siblings). However, nothing else (perhaps most importantly, their own sexual orientation) is known about these siblings.

This is what we meant with "shift in the level of analysis".

As for the violation of the assumption of independence, we needed to reconsider and now think that it is actually plausible to assume that the siblings reported by participants can be treated as independent observations with respect to the variable "older brother" (1 = yes, 0 = no). If we could say that knowing that the jth sibling reported by the ith study participant is an older brother (or not) changes our probability assignment of the event that the $j + 1$th sibling of the ith participant is also an older brother, then we could make a case for that the siblings are (stochastically) dependent. However, we do not see how such a dependence could be justified.

We thus removed all instances in the manuscript where the possibility that observations are not independent is raised.

@MROB: This is how we understand the claims made by, e.g., Blanchard (2014).

@OR: No, the OR makes the same shift in the analysis level as the OBOR. The OR would be a better alternative to the OBOR.

Line 374: I believe there's also a substantive reason why the number of younger siblings should not be controlled for: it may be an outcome of the target's sexuality, so controlling for it might induce collider bias (e.g., between the number of older brothers and sexuality, which may both affect parents' decision to have more children).

Response: We are not sure if we can follow. Our understanding of the term "collider bias" in this context is as follows: the number of older brothers and sexual orientation both show no direct association with each other, however they independently affect parents' decision to have more children (?). And by adjusting for the number of younger siblings, the number of older brothers and sexual orientation show some degree of association. This sounds interesting, but we are not sure how to put the pieces together here, given the lack of availability of such data structures in this research literature.

Reviewer 3: Your understanding is basically correct, although it need not be the case that there is no causal association Older brothers -> Homosexuality. A collider bias would arise when we condition on a variable that is causally affected by e.g. outcome and predictor in a regression (e.g. Homosexuality -> more younger siblings, More older siblings -> More younger siblings). It seems like more individual-level data would be great for this literature. And it appears Ablaza et al. 2022 conditioned on the number of younger siblings. You might want to point out this problem before their method gets enshrined as the right way.

Line 378: Your discussion makes it sound like what Gelman and Stern propose does not introduce collinearity, but simply controlling for older siblings does. But…the models will be exactly the same, model fit will be the same. The coefficients will just have a different meaning. But if you use the model to derive average marginal effects or something, the answers will be precisely the same. There's three possible numbers that can be included in a model: #OS, #OB, #Older. The latter is the sum of the former two. Now if you do some linear combination of the first two and combine it with the latter that is…still the same pieces of information, you're just splitting things up in a different way. The coefficients will have a different interpretation. But the model implied effects will be the same, as the model makes exactly the same prediction for everyone. Here's a quick simulation:

```
set.seed(12345)
n <- 10000
older_sis <- rbinom(n, size = 4, prob = .50)
older_bro <- rbinom(n, size = 4, prob = .50)
older <- older_sis + older_bro
# Some arbitrary older brother effect
hom_prob <- 0.2 + older_bro * 0.10 + rnorm(n, sd = .1)
# using a probit link here
hom <- hom_prob > .5
table(hom)
```

```
## hom
## FALSE  TRUE
##  7637  2363
```

```
# using an inadequate lpm because details dont matter here
m1 <- lm(hom ~ older_bro + older)
m2 <- lm(hom ~ older_bro + older_sis)
# lets do what gelman and stern suggest
diff <- older_bro - older_sis
m3 <- lm(hom ~ diff + older)
p1 <- predict(m1)
p2 <- predict(m1)
p3 <- predict(m1)
# it is all the same
p1[1:5]
```

```
##         1         2         3         4         5
## 0.0140099 0.4538977 0.4538977 0.6738416 0.2390805
```

```
p2[1:5]
```

```
##         1         2         3         4         5
## 0.0140099 0.4538977 0.4538977 0.6738416 0.2390805
```

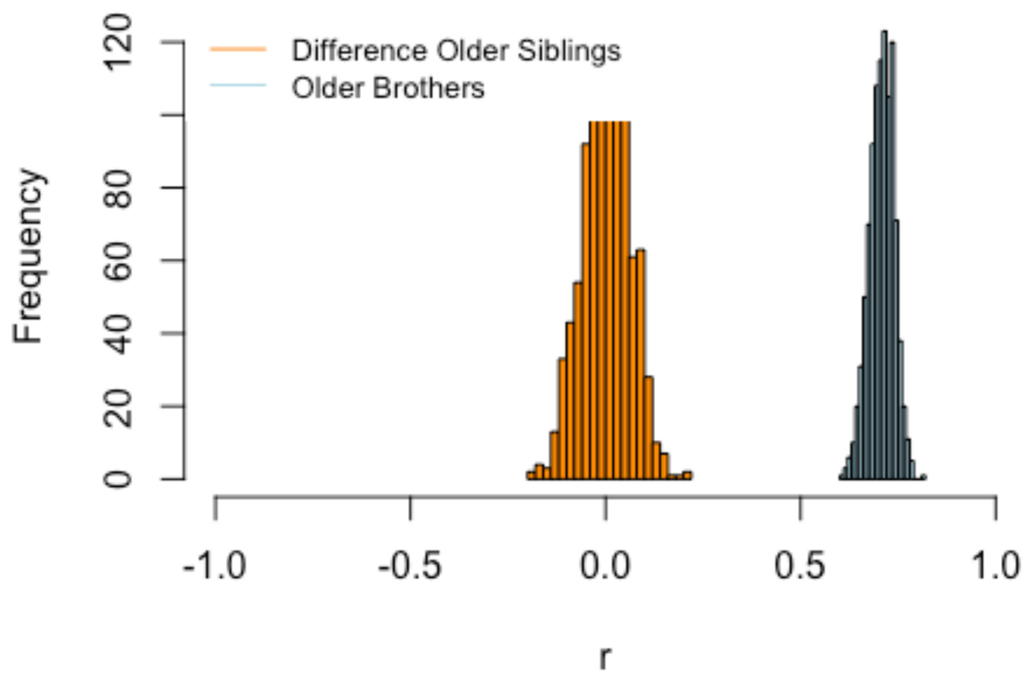
```
p3[1:5]
```

```
##         1         2         3         4         5
## 0.0140099 0.4538977 0.4538977 0.6738416 0.2390805
```

Response: Yes, the model seems to introduce no collinearity as:

```
set.seed(NULL)
n <- 200
sim_corr <- vapply(1:1000, function(x) {
  older_sis <- rbinom(n, size = 4, prob = .50)
  older_bro <- rbinom(n, size = 4, prob = .50)
  older <- older_sis + older_bro
  c(m1 = cor(older_bro, older), m2 = cor(older_bro-older_sis, older))}, double(2))
h1 <- hist(sim_corr[1 ,], plot = F, breaks = 20)
h2 <- hist(sim_corr[2 ,], plot = F, breaks = 20)

plot(h1, col = "lightblue", xlim = c(-1,1), xlab = "r", main = "")
plot(h2, col = "darkorange", add = T)
legend("topleft", legend = c("Difference Older Siblings", "Older Brothers"),
    col = c("darkorange", "lightblue"),
    lty = c(1,1),
    box.lty=0,
    cex = 0.8)
```



We did not intend to communicate that the correlation between the number of older brothers and the number of older siblings should be regarded as an issue. That is, we do not think that collinearity is a problem in any way here. The note about collinearity in the manuscript was irrelevant to the issue at hand. We thus deleted it in ms. revision. The "advantage" of the second model is that it gives us a direct estimate of the effect of interest.

As a side note: when there is a general birth order effect, m1 will give you no significant coefficient for older brothers. M2 will give you a significant coefficient for both older brothers and older sisters (which is, from a causal perspective, the right answer – if there is a birth order effect, that will be mediated through both older sisters and older brothers). M2 will give you no significant coefficient for the differences variable. See code below:

```
hom_prob2 <- 0.2 + older * 0.10 + rnorm(n, sd = .1)
hom2 <- hom_prob2 > .5
m1 <- lm(hom2 ~ older_bro + older)
m2 <- lm(hom2 ~ older_bro + older_sis)
m3 <- lm(hom2 ~ diff + older)
summary(m1)
```

```
##
## Call:
## lm(formula = hom2 ~ older_bro + older)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94758 -0.32250  0.05957  0.24881  0.82370
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.012942   0.010329  -1.253    0.210
## older_bro    0.003575   0.004766   0.750    0.453
## older        0.189245   0.003430  55.178   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3416 on 9997 degrees of freedom
## Multiple R-squared:  0.3795, Adjusted R-squared:  0.3794
## F-statistic:  3057 on 2 and 9997 DF,  p-value: < 2.2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = hom2 ~ older_bro + older_sis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94758 -0.32250  0.05957  0.24881  0.82370
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.012942   0.010329  -1.253     0.21
## older_bro    0.192820   0.003397  56.769   <2e-16 ***
## older_sis    0.189245   0.003430  55.178   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3416 on 9997 degrees of freedom
## Multiple R-squared:  0.3795, Adjusted R-squared:  0.3794
## F-statistic:  3057 on 2 and 9997 DF,  p-value: < 2.2e-16
```

```
summary(m3)
```

```
##
## Call:
## lm(formula = hom2 ~ diff + older)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94758 -0.32250  0.05957  0.24881  0.82370
##
```

Response: Yes, this is precisely the reason why the model which includes the difference and sum as predictors should be preferred over the model which includes the number of older brothers and the number of older sisters, if one is interested in the highly specific claim that the number of older brothers, but not older sisters, are associated with homosexual orientation (see Gelman and Stern, 2006, who seem to make exactly the same point). Whether older siblings in general contributed to an individual's sexual orientation would surely be an interesting question in its own right, but it is of no concern here. Our perspective is this: We are first and foremost interested in the claims about how to assess the key observation of the FBOE made by Blanchard et al., regardless of the true underlying causal mechanism or what other patterns may be observable in the available data. We apologize if our manuscript came off as making causal claims, or as if we were interested in the substantive theory of the FBOE. This is certainly due to a lack of precise communication on our part in the originally submitted ms. Also, in our opinion m1 and m3 clearly allow assessing whether the association between older brothers and sexual orientation is stronger than the association between older siblings and sexual orientation, which is what the FBOE predicts.

We corrected and revised the Introduction section of our ms. accordingly, now clearly stating that the focus of this work is on the technical aspects of proposed methods, given a certain set of assumptions put forth by Blanchard et al.

Beginning of Part 2: Your explanation of what you mean by false-positive rate is a bit contorted. I guess the distinction your try to explain is "variance false positives" (that is what the statistical literature is mostly concerned with) vs. "bias false positives" (that is what, for example, the causal inference literature is often concerned with – Per Engzell used these terms in a paper I wrote with him). Maybe you just want to talk about "rate of mistaken conclusions" or something like that, that is much easier.

Response: Yes, and thank you for the recommendation. However, we are not convinced that the terms used in Engzell and Rohrer describe exactly what we meant by false-positive rates. The problem with most of the models presented here is that they are misspecified in that the statistical model corresponds only poorly to the substantive model (which does not have to be causal), and hence the widely applied heuristic of equating statistical hypothesis tests with substantive hypothesis tests would lead to a higher frequency of false decisions pertaining to the substantive hypothesis. We now give a longer description of the reasoning underlying the performance criteria we used in the simulation study.

Methods of Part 2: I find it a bit confusing that you simulate older brother effects, but then also simulate independent effects of the number of older siblings (by setting the proportion of the number of older siblings differently for the different samples). What is an effect of the number of older siblings independently of an effect of older brothers? An effect of older sisters?

Response: Thank you very much for this excellent point, which we never thought of. The proportions of older siblings could differ between groups for many reasons, including non-random or non-representative sampling or a genuine tendency for homosexual men to have more older siblings than heterosexual men. It makes sense to consider these effects as independent, should someone suspect that the effect of older brothers on sexual orientation is stronger than that of older sisters (as proponents of the FBOE do).

Maybe you could try to setup your description of your simulations along a data generating process. So, you simulate a number of older brothers and a number of older siblings. Together, these are the number of older siblings. Now you generate the odds of homosexuality as a function of these variables. Yes, this requires you to reason backwards – deduce the birth order effect from the numbers reported in the literature – but your train of thought will probably be much easier to communicate.

Response: Thank you for your suggestion of describing the data-generating process. This would run thus:

> Let OB, OS, YB, YS denote the random variables "number of older brothers", "number of older sisters", "number of younger brothers" and "number of younger sisters", respectively. Let $\mathbf{X} = (OB, OS, YB, YS)$ be a random vector with correlation matrix $\mathbf{R}$. Let n be the sample size. We assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{iid}}{\sim} \text{Pois}_{\mathbf{R}}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_{OB}, \lambda_{OS}, \lambda_{YB}, \lambda_{YS})$ is the vector of rate parameters. Furthermore, let Y_i, $i \in \{1, \ldots, n\}$ be the sexual orientation of the ith participant (0 = heterosexual, 1 = homosexual), $Y_i \sim \text{Bernoulli}(p_i)$, where the probability of homosexual orientation p_i is given by $p_i = \text{logit}^{-1}(\text{logit}(0.02) + \ln(1.33) \times OB_i)$.

Line 453: why would there be a positive correlation between the number of older brothers and the number of older sisters? Is that really an empirical observation? Like just some general fertility confounding or something? Also, what exactly would such a correlation induce in which model? You make it sound like this would necessarily result in an older sister effect that is non-zero, but if both types of siblings are in the same model, this shouldn't happen, should it? The inclusion of older brothers blocks the backdoor path. You would just find a zero-order association.

Response: The origins of the positive correlation among older (and also younger) siblings initially caused some confusion among us as well. We believe that there is an analytic proof, unfortunately, we were not able to come up with one. We would like to emphasize two points here:

First, the positive correlation between older siblings is a claim made by Blanchard et al. and hence we adopted it as an assumption. So it is Blanchard et al. who assert that there is a necessary older sisters effect due to this correlation ("effect" referring to the statistical association between sexual orientation and the number of older sisters).

Second, Blanchard et al. never gave any explicit or clear description/explanation as to how this correlation would be relevant to their statistical methods. In our opinion, the small correlation we implemented in our simulation (and observed in real-world data) does nothing to the models. It is also Blanchard et al. (e.g., 2014), who have stated that homosexual men have more older brothers on average than heterosexual men and that - since the number of older brothers and older sisters is positively correlated - they also have (slightly) more older sisters on average than heterosexual men.

We also wrote a simulation (mostly to convince ourselves that the positive correlation makes sense): If we assume (for convenience) that the number of total children in a family are distributed according to a Poisson distribution (but a negative binomial or any other sensible discrete distribution would do as well), and that each sibling is independently either male or female, the positive correlation can be observed:

Denote by X_j the random variable "number of children in the jth family", where $j \in \{1, \dots, N\}$. We assume that all of the X_j are independently and identically distributed according to a Poisson distribution with rate parameter λ (i.e., the average number of children in a family). That is $X_j \overset{\text{iid}}{\sim} \text{Pois}(\lambda)$. The Poisson distribution is chosen purely due to its convenience, in essence, the only purpose it serves is to generate integer values. We set $\lambda = 2$ (but any other mean number of children will do), $N = 10^4$.

```
x <- rpois(10^4, 2)
x <- x[x > 0] # Remove 0 counts
```

Next, for each "family" in `x`, we draw as many Bernoulli random variables as there are children in `x` to determine the sex (male or female). For instance, if the jth family in `x` has 3 children, we simulate one Bernoulli random variable for each child. That is, let Y_{ij} denote the sex of the ith child in family j, we assume that $Y_{ij} \overset{\text{iid}}{\sim} Bern(p)$, where p is the probability of "male", which seems to be .515 (rather than .5). We can then deduce the number of each sibling type for each of the children. We wrote a function that computes these numbers:

```r
sib_comp <- function(no_children){
  # First, consider families with more than one child
  if(no_children > 1){

    # Create empty array for storing sibship composition
    temp <- array(0, c(no_children, 5),
            dimnames = list(rep("", no_children),
                    c("sex", "ob", "os", "yb", "ys"))
    )

    # Randomly generate sex of child (birth order is implied by the row
    # number in which the child appears)
    temp[, "sex"] <- rbinom(no_children, 1, .515)

    # Use the information in "sex" to determine number of each sibling type
    temp[2:no_children, "ob"] <- cumsum(temp[1:(no_children-1), "sex"])
    temp[2:no_children, "os"] <- cumsum(!temp[1:(no_children-1), "sex"])
    temp[1:(no_children-1), "yb"] <- rev(cumsum(temp[rev(2:no_children), "sex"]))
    temp[1:(no_children-1), "ys"] <- rev(cumsum(!temp[rev(2:no_children), "sex"]))

  } else {
     temp = cbind(c(rbinom(1,1,.515), 0, 0, 0, 0))
  }
  return(temp)
}
```

For instance, we can simulate the sibship "composition" of a family with 5 children by calling

```r
sib_comp(5)

##  sex ob os yb ys
##    1  0  0  4  0
##    1  1  0  3  0
##    1  2  0  2  0
##    1  3  0  1  0
##    1  4  0  0  0
```

If we apply this function to each of the N "families" and randomly sample one child per family, the correlation can be observed. Here's a function that does this

```
draw_sample <- function(fam_size){
   smpl <- array(0, c(length(fam_size), 5),
        dimnames = list(rep("", length(fam_size)),
             c("sex", "ob", "os", "yb", "ys"))
   )
   for(i in 1:nrow(smpl)){
    temp2 <- sib_comp(fam_size[i])
    index <- sample(1:nrow(temp2), 1) # Sample person in family
    smpl[i ,] <- temp2[index ,]
   }
   return(smpl)
}

sim_sample <- draw_sample(x)

head(sim_sample)

##  sex ob os yb ys
##    0  0  0  0  0
##    1  0  1  0  0
##    0  0  1  0  0
##    0  1  0  0  0
##    1  1  1  1  1
##    0  0  0  0  0
```

The resulting correlation matrix is:

```
cor(sim_sample[, c("ob", "os", "yb", "ys")])

##             ob          os          yb          ys
## ob  1.00000000  0.18709946 -0.03366689 -0.02404879
## os  0.18709946  1.00000000 -0.03734614 -0.02794859
## yb -0.03366689 -0.03734614  1.00000000  0.18162268
## ys -0.02404879 -0.02794859  0.18162268  1.00000000
```

Altogether, the correlation for older brothers and older sisters is similar in magnitude to the one we report in Table 3.

Figures: I'd suggest that for Figure 1, you include clearer labels (in the actual figure) so that the figure can "stand on its own." Right now, it's impossible to make sense of the figure without reading the corresponding paragraph in parallel. I'd also really try to make the figure cognitively ergonomic. Like, where would values need to fall for unbiased estimation? There is no indication for that in the figure. I think this might require you to rethink the figures and maybe replot from scratch. However, I believe that this may be worth the effort. This is a very central part of your manuscript, and I'd want to make sure that it is accessible.

Response: Thanks for pointing this out to us: we see that the figure descriptions, as originally submitted, were insufficient, and we have made appropriate amendments. We would also have liked to add “unbiased” estimates, but the problem is that here the only reasonable meaning of the term “bias” is in the statistical sense of the bias of an estimator for a population parameter. The estimates shown in Figure 1 are all unbiased in this sense; they provide an unbiased estimate for an effect that does not correspond to the effect of interest. At any rate, we made efforts to make the figure more legible, by using additional labels and colors.

Minor Comments
Abstract, “For a quarter of a century researchers investigating the origins of sexual orientation have largely ascribed to the fraternal birth order effect (FBOE) as a fact, holding that older brothers increase the odds of homosexual orientation among men through an immunoreactivity process.”: I’m not a native speaker, but I’m not sure whether this sentence works in English. In any case, it is a bit convoluted. Maybe you could simplify this a bit for increased accessibility (e.g., start with what the fraternal birth order effect is, then talk about its status in the literature).

Abstract, “This yielded highly inconsistent and moreover similarly sized effects across 64 male and 17 female samples (N = 2,778,998), compatible with an excess as well as with a lack of older brothers in both groups, thus, suggesting that almost no variation in the number of older brothers in men is attributable to sexual orientation.”: There is a good chance that at this point, readers won’t know why the “similarly sized effects” are relevant. Once again, I think you can break this down into multiple sentence to make it more accessible.

Response: Thank you for these additional suggestions, highly appreciated. We have rewritten the Abstract, emphasizing that “similarly sized effect” refers to the difference in the effect between male and female samples.

Line 50, “Moreover, the FBOE implies an observed excess of older brothers in homosexual men, whereas the reverse does not follow logically.”: I don’t think I follow. If there is a FBOE, then homosexual men will have more older brothers; but if older men have more older brothers, the FBOE does not follow? You are right that confounding third variables might result in a spurious association between older brothers and homosexual orientation. But confounding factors could also hide a true causal effect (say, if there is a confounder that causes an excess of older sisters). So, an FBOE does not necessarily imply an observed excess of older brothers. (I don’t think it’s plausible that a true FBOE might be hidden because of some sort of confounding, but if you’re trying to make a logical statement, it should be waterproof in my opinion)

Response: You are right, this statement in our originally submitted ms. did not really make sense. Rather, we meant that the FBOE, as formulated by Blanchard et al., necessitates the observed excess of older brothers relative to older sisters in homosexual men, but that an observed excess of older brothers relative to older sisters in homosexual men (conditional on all relevant variables being adjusted for) would not be sufficient in supporting the FBOE and the biological explanation offered by Blanchard. For the sake of clarity of our presentation, we have rewritten this paragraph in its entirety.

We did not allude to confounding or causal factors, but rather to the crud factor (everything is correlated with everything for whatever reasons, we also added this to the discussion). Whether the FBOE necessitates an excess of older brothers or not depends on the background assumptions made. As you point out (and we agree very much), there may be confounding variables that could hide or distort this excess. We were not clear about this, but Blanchard et al.'s formulation of the FBOE necessitates the excess of older brothers. This is the framework we adopt; we assume the context laid out by Blanchard et al. as given.

We have made various efforts to be more precise in the revision of the Introduction section.

Line 69: What is a "smaller excess" of older sisters? This section was really confusing to me, maybe because it is so unclear how this statement relates to the previous statements. I think it might be worth trying to rewrite this section for greater clarity (I'd go for something like: second, the FBOE is often combined with the claim that homosexual men also have a slightly higher number of older sisters. This is borne out of empirical observation…and explained the following way…[which needs to be spelled out a bit more, in my opinion – there's a step between miscarriages of XY male foetuses and an excess of older sisters that should be spelled out]).

Response: Yes, this was confusing and the terms "excess" did not make sense here (namely, excess relative to what? - thanks for pointing this out to us). Accordingly, we have rephrased this text passage.

Line 116: "These observations combined would strip the FBOE and the MIH of its verisimilitude": In general, I liked this section because you spelled out your argument quite clearly. But this conclusion seemed a bit anti-climactic (and I don't know whether epistemiologically, it makes sense to say that something is "stripped of its verisimilitude"). Maybe you could make it a bit more straightforward (think "These observation would undermine both the FBOE and the MIH as an explanation for homosexual orientation in men.")

Response: Once more, thanks for this further attention to detail and usage of terms. It may be that the word "verisimilitude" was misplaced, or that our understanding of the word was quite off from what it actually means. Either way, we have rephrased this text passage, also incorporating the comments by Reviewer 3 (Ruben Arslan).

I found the phrasing "implicitly defined at the level of siblings" a bit confusing. You clarify later, so it is not a big issue, but I certainly stumbled over it.

Response: Agreed, we removed it, as we mention it later anyways.

Reviewer 2 (Anonymous)

Additional Comments

This paper assesses the empirical foundations of the Fraternal Birth Order Effect (FBOE) from three separate perspectives: (a) a mathematical assessment of the relevant dependent variable (b) a simulation study (c) a multiverse meta-analysis.

Before discussing each of these in turn, it is important to note that the authors rightly differentiate between (i) The FBOE as a causal statement (i.e., older brothers increase the likelihood of homosexual orientation) (ii) The FBOE as a merely correlational or observation statement.

Within the FBOE as a causal statement, it seems important to further distinguish between (i-weak) Older brothers increase the likelihood of homosexual orientation of younger-born males (i-strong) Older brothers increase the likelihood of homosexual orientation younger-born males, but older sisters, younger brothers, and younger sisters have no effect on it.

Moreover, an even more specific claim than (i-strong) is (mih i-strong) The FBOE is a result of the maternal immune hypothesis (MIH) because other things follow from this (e.g., there should be no association between number of older brothers and likelihood of homosexuality in women).

I believe the authors are specifically interested in (mih i-strong). At least to my reading, they make this pretty clear in their Introduction and it is implied by their discussion of the association between orientation and older brothers in females in the third part of their paper.

However, sometimes they do not stay carefully attuned to this throughout the paper. I think they need to be more clear about what version of the FBOE they are specifically interested and how various analyses they do follow from or depend on the version of interest. As of now, I think this is too loose and the paper risks overclaiming (i.e., because it is attacking a narrower position than sometimes appears to be the case).

Response: Thanks, this is a very helpful comment! As we detailed in our responses to Reviewer 1, we made appropriate changes throughout the manuscript, such that we now clearly define the conditions (or framework) we assume. We are concerned with how one can answer the necessary observation (as asserted by Blanchard et al.) of a higher number of older brothers relative to older sisters from a data-analytic perspective. We are not concerned with substantive claims such as the origins or causal factors of sexual orientation. Also, we were not able to verify the distinction between i-weak, i-strong and mih-i-strong. The literature on the FBOE and the claims by Blanchard et al. clearly state that older brothers but not older sisters increase the odds of homosexual orientation via the MIH. This is the framework we adopt.

While on the topic of overclaiming, the authors’ three-perspective triangulation approach is nice and appreciated and impressive.

Response: Thank you for this positive evaluation.

However, fundamentally, all three perspectives are mathematical / statistical in nature and so are in a sense three variations on a single perspective rather than three separate perspectives. To put the nail in the coffin of the FBOE (or rather some particular version of the FBOE like mih i-strong) as I believe the authors seek to do requires a more holistic assessment that goes beyond the mere mathematical / statistical and engaging with and triangulating across various lines of evidence (e.g., evidence of antibodies as discussed by the authors with respect to the MIH).

Response: We consider solely the technical data-analytic context. Hence, these approaches can be considered as separate. Our aim was not "to put the nail in the coffin" of the FBOE, but rather to demonstrate that most of the evidence for the FBOE is based on inadequate/ambiguous (or even faulty) statistical reasoning.

Consequently, I think the title of the paper is a strong overclaim. The manuscript should indeed cause those who view the FBOE as "settled science" to question that view and remain more tentative (although maybe they should not have had such certainty in the first place). However, the manuscript does not seem to have unequivocally refuted the FBOE.

Because the title and some of the writing tends to read like the authors believe they have, I would ask them to be carefully attuned to avoiding that in any revision. In doing so, it is important to emphasize that even a proof that (mih i-strong) is false (which is lacking) need not imply (i-strong) or (i-weak) is false. In fact, I think there may be nothing in the paper speaking to either (i-weak) or (ii), which is of course okay! It is just important for the authors to make clear to readers that there are various version of the FBOE and they might each have differential support in light of the arguments put forth in this paper.

Response: As for the ms. title: given that (almost) all of the statistical evidence in favor of the core observation of the FBOE is based on statistical analyses, which are ambiguous, if not faulty, previous claims about the magnitude of the association between older brothers and sexual orientation (the theoretical estimand) are best described as statistical artefacts.

To our knowledge Blanchard et al. put forth only one version, namely the one relying on the MIH. Conditional on this model, Blanchard et al. derive statements about observable variables (homosexual men have more older brothers relative to older sisters than heterosexual men). Moreover, Blanchard et al. prescribed a set of methods for assessing these observable variables statistically. Our work is concerned with these methods, not with the "true" mechanism (i-strong, i-weak, mih-strong, or any other version of the FBOE).

I now discuss each of the three perspectives discussed in the paper.

A. Mathematical assessment of the relevant dependent variable

Before proceeding, I have an item of confusion. On Lines 167-169, the authors write "P(OB) should be read as 'the probability that a given sibling is an older brother to the participant who reported him or her.'" This seems underspecified. I think we would need subscript i (denoting the participant who reported) and a subscript j (denoting the given sibling). Even if we take the focal participant as given (so that the subscript i is understood), I do not see how we avoid the subscript j. Without it, P(OB) instead would seem to denote the incidence of having an older brother at all rather than that some specific sibling is an older brother.

This is important because many of the key quantities discussed in the paper such as the older brothers odds ratio (OBOR; Equation 1) follow from P(OB) or some conditional version of it and therefore cannot be interpreted without this issue clarified. Can the authors clarify?

Response: From the above context, it seems that incidence means proportion, which is an estimator of the probability of an event. Then P(OB) correctly refers to the probability that a given sibling out of all siblings reported by a group of participants is an older brother. The information that a sibling came from participant i is irrelevant to the computation of the OBOR as formulated by Blanchard et al., so adding indices could wrongly leave the impression that this information might be important. For the MROB and MPOB on the other hand, indices are relevant, which is why we added indices in the corresponding equations.

Moving on, I agree with the decomposition of the older brothers odds ratio (OBOR) in Equation 4 and the comments immediately below that in Lines 217-222 about the logical implications that follow from it. However, what I am less clear about is the comments in Lines 222-223 that the FBOE requires "the homosexual group must show a greater proportion of older brothers among older siblings." Does it? The quotation of Blanchard (2018a) on Lines 45-47

"Older brothers increase the odds of homosexuality of later-born males, whereas older sisters, younger brothers and youngers sisters have no effect on those odds."

seems very clearly to be a statement about numbers and not one about proportions. Therefore, it does not seem the FBOE requires this as claimed by the authors.

Response: Blanchard et al. chose to operationalize this definition (via the OBOR) and thus (a) to regard the siblings reported by the participants as the sample, and (b) to analyze the proportions of older brothers, older sisters, etc. This is implied by using an odds ratio, which is a quantitative statement about the odds of an event occurring vs. not (e.g., Agresti, 2013, pp. 44-45) and is sensibly estimated using proportions. Also, proportions and counts contain exactly the same information, as one can easily be transformed into the other.

According to Blanchard et al. and the MIH, a greater number (and hence proportion) of older brothers relative to older sisters among homosexual men is exactly what the FBOE requires. Suppose we have two groups, one consisting of homosexual men, the other of heterosexual men. Each man in the homosexual group and each man in the heterosexual group has exactly k older siblings. Then the FBOE as a statement about numbers implies that the average number of older brothers in the homosexual group is greater than the average number of older brothers in the heterosexual group. For the average in one group to be greater than the average in the other group, there must be more older brothers in one group than in the other group, and the number of older siblings in each group must add up to $k \times n$ (where n is the group size). Then, the proportion of older brothers in both groups is given by $ob/(k \times n)$, where ob denotes the number of all older brothers reported by the group. If ob is larger in one group vs. the other, $ob/(k \times n)$ is also larger in that group versus the other.

Regardless of whether or not the FBOE requires this, the authors' solution is to eschew the OBOR (as well as the OBOR and older sisters odds ratio (OSOR) in tandem) and instead to consider the odds ratio (OR; Equation 6), namely the ratio of older brothers to older sisters among homosexuals divided by that same ratio among heterosexuals.

However, it is not clear this follows from (i-strong) or (mih i-strong). I would think either the incidence of having an older brother among homosexuals to heterosexuals OR the ratio of the number of older brothers among homosexuals to heterosexuals would be adhere more closely to (i-strong) and (mih i-strong).

That is, these claims seem to be more about the incidence and numbers of older brothers and not about them as a fraction of older siblings; again, see the quotation of Blanchard (2018a) on Lines 45-47 (as an aside, mih i-strong arguably suggests the greater the number of older brothers, the higher the likelihood of homosexuality not just that incidence of older brothers increase the likelihood).

Response: Yes, the FBOE states that each older brother (not just *a* older brother) increases the odds of homosexual orientation by a certain amount. But we were not trying to estimate or assess this effect per older brother (as such data are not available). Moreover, neither the OR nor the OBOR are used to estimate the extent to which having one or more older brothers increases the odds of homosexual orientation. Instead, they estimate the extent to which being reported by a homosexual versus heterosexual participants increases the odds of being an older brother. Importantly, quantifying the FBOE by assessing the difference in the odds of being an older brother of a homosexual versus heterosexual male was Blanchard et al.'s idea, not ours. We merely state that if one considers their approach to be valid, then the OR of older brothers vs. older sisters would be a better choice of effect size than the OBOR.

What you seem to suggest in the second part is this:

$$\frac{\text{Number of Older Brothers reported by Hom}}{\text{Number of Older Brothers reported by Het}}$$

We can multiply this by $\frac{1}{N}/\frac{1}{N}$, where N is the total number of all siblings

$$\frac{(\text{Number of Older Brothers reported by Hom})/N}{(\text{Number of Older Brothers reported by Het})/N}$$

So, now the numerator and denominator consist of the proportions of older brothers who were reported by homosexual and heterosexual participants respectively. We write this as

$$\frac{Pr(OB, Hom)}{Pr(OB, Het)}.$$

This is equivalent to (by the definition of conditional probabilities)

$$\frac{Pr(OB|Hom)Pr(Hom)}{Pr(OB|Het)Pr(Het)}.$$

Not only does this ratio suffer from the same problem as the OBOR (we could decompose Pr(OB|Hom), Pr(OB|Het) again and show that it is affected by the proportion or, equivalently, the number of older siblings in either group in the first place), but it is also affected by the proportion of homosexual individuals in the sample.

Usually, samples have many more heterosexual individuals than homosexual ones, which means that most older brothers reported in a sample will be reported by heterosexuals. So unless the two groups are equal in size or there are more homosexual than heterosexual individuals, the ratio of counts you suggest will almost always be less than unity.

Thus, while I agree with the authors' criticism of the OBOR, I do not think their proposed OR solution follows. It seems to involve a change in the very nature of the FBOE from being a statement about incidence or numbers to one about proportions. These are very different things.

Response: Analyzing count data in terms of porportions is quite adequate and a standard approach for analyzing binary outcomes (Agresti, 2013).

Now, I understand the whole use of OBOR, OR, etc. in the first place are motivated by (a) the idea that homosexuals and heterosexuals might have different family sizes (and indeed seem to empirically, perhaps as an artefact of sampling/measurement) along with the idea that (b) difference in family size arguably should be adjusted for in some manner. Maybe differences in family size do need adjusting.

Response: Yes, this is an important substantive point, and we agree. There are numerous possible explanatory variables lurking at the family-level that conceivably are potential confounders. If one would like to make substantive claims about sexual orientation, then considering data at the family level and all kinds of other possible explanatory variables would be a necessity. We added this point to the discussion.

However, it seems to me that either or both of the following are or could be the case (i) the best way to adjust would be to obtain better samples/measures (perhaps at the level of the entire family) rather than to rely on statistical or mathematical adjustments like both Blanchard and the present authors do (ii) perhaps differences in family sizes are not a mere artefact of sampling/measurement but require an extension of the underlying theoretical model.

Response:

@(i) This is not what we sought to do, and we did not write anything along these lines. We checked the manuscript carefully and did not find any instance where we make substantive claims about how older brothers contribute to sexual orientation, and that this could be achieved

by statistical adjustment/controlling.
We showed that the OBOR does not address the question asked by Blanchard et al., and that for that specific question (is there an excess of older brothers, but not sisters, among homosexual men, given that there are no other confounders?) the OR is better suited as it fixes the OBOR, if one insists on analyzing the data as Blanchard et al. did and adopts their assumptions about relevant confounders.

@(ii) This is possible, but we are not primarily concerned with the substantive aspect of Blanchard et al.'s claims. We are only concerned with the technical claims made by Blanchard et al. pertaining to the statistical evaluation of the core observations of the FBOE.

Lines 342-344: I think this overstates matters. The statistical model does not say they have no effect but instead says they have an effect but in this very tightly specified and inflexible manner. Please be precise.

Response: Thank you for pointing this out to us. By "no effect" we meant that by excluding the lower-level terms in the statistical model, their regression coefficients (partial effects) are set to zero. We added this information in ms. revision. However, the statistical model does imply that the constituent variables making up the interaction have no effect on their own, but only some undiscernible interplay between the variables related to sexual orientation (Kronmal, 1993).

More minor comments
Lines 378-379: It is wrong to say "Gelman and Stern (2006) proposed the difference between #OB and #OS as one predictor and #Older as a second predictor." Gelman and Stern never discuss older brothers, older sisters, etc. in their paper. I know what you are trying to say here but you need to say it in a manner that is precise and correct rather than attribute something specific to Gelman and Stern that they did not actually say.

I note as an aside that a perhaps useful additional approach would also be to consider a regression on OB and OS with coefficients forced to be equal versus allowed to vary. One could perhaps also look at OB alone.

Response: We took this matter seriously and contacted Prof. Gelman, who kindly confirmed via personal communication that this was exactly the model they suggested. The model and discussion about older brothers, older sisters, etc can be found in the final paragraph of section 3 on page 330 in Gelman & Stern (2006).

B. Simulation Study

I found this section of the paper poorly motivated and therefore not all that compelling. It also seemed like a bit much and rather unnecessary after the prior section. Further, whereas the link between the first and third parts of the paper was quite strong (i.e., an argument for OR over OBOR suggests revisiting prior meta-analyses but using OR in place of OBOR), this section just seemed to dangle in the middle. I think it could profitably be cut, especially given the length of the paper.

Response: Remarks on the simulation study made by both Reviewers 2 and 3 were quite positive. They instead suggested that we cut the mathematical assessment in Part I. We decided to leave it this way. To some readers, the simulation study might be more accessible. In addition, we rewrote much of this section, emphasizing our motivation for conducting this simulation.

That said, I do believe the simulation description was clear (although choices of parameters, distributions, sample sizes, etc. were not always well justified). Also clear was the description of the models used.

Response: Thanks! Yes, these choices were made from convenience. We also published and documented the analysis code in the supplementary materials on the osf website. This way, readers can play around with the parameters and simulate different scenarios.

Some final questions and comments: why do you employ a multivariate Poisson in this simulation? Do family sizes tend to be distributed in this manner? Often real data is overdispersed and so an overdispersed-Poisson or the Negative Binomial distribution might be a better choice. And is it reasonable to assume the same correlation structure for families of homosexuals and heterosexuals?

Response: The Poisson distribution was chosen out of convenience. We needed a way of generating count data. Whether human family size actually is distributed according to a Possion distribution is not really relevant at this point. We did not try to model real-world compositions of families, but rather tried to demonstrate the poor properties of the methods suggested to model data which occur in counts. It may well be that the negative binomial distribution would yield a better approximation, but, in this case, it really is a Poisson distribution with an added dispersion parameter. Estimating this dispersion parameter from summary data is difficult and often returns impossible values (addressing this issue would increase the length of this manuscript considerably). On the other hand, little seems to be gained, as the additional variance incurred by using the negative binomial distribution, compared to that of a Poisson distribution.

Concerning the equal correlations in both groups: The claim that the positive correlation between older brothers and sisters (regardless of whether we look at homosexual or heterosexual men) plays a role was made by Blanchard et al. They never provided any indication as to how or whether this correlation should be different in the two groups. Given the little information available, equal correlation structures appeared to be the most reasonable assumption to make. We added this information to the description.

Further, the evaluation of the simulation focused on two-sided significance tests. I would much rather see an evaluation based on how well these approaches are able to estimate theta (bias, RMSE, CI coverage, CI width, etc.), which you say on Line 538 is what they are designed to do. Nonetheless, even remaining within the significance testing framework, the two-sided test approach seems very odd given that the FBOE predicts a direction. It would therefore make more sense to use one-sided tests.

Response: Agreed on these points. However, we clearly stated that only some of the models provide estimates for theta. We would much rather report these more informative statistics about how well the approaches estimate theta. But most of the approaches put forth by Blanchard et al. do not provide estimates for theta (the OBOR, the logistic regression models using the MROB and MPOB) – these models were clearly invented with the intention to subject them to hypothesis tests and argue based on statistical significance. These models are misspecifications and therefore do not estimate theta, but rather something else. They are not biased, in the sense that their expected values being equal to the population parameter on average. They simply do not answer the question (the model does not address theoretical estimand). We chose to evaluate the methods in the simulation via the NHST framework for two reasons: First this is the dominant decision rule employed by Blanchard et al. All substantive claims about older brothers increasing the odds of homosexual orientation are based on statistical significance. When claiming that older brothers increase these odds based on significance tests of falsely specified models, Blanchard et al. are making inferences about theta being 0 or not. We show that their decision rule does not have the frequency property of falsely rejecting in only 5% of cases (conditional on all other assumptions being met). This also explains why we use two-sided tests: because Blanchard et al. mostly employed two-sided tests.

Line 425: theta is incorrectly written as cursive theta here.

Response: Corrected in the revised ms.

Line 440-443: Can you explain why you let pi vary for the homosexual group but not the heterosexual one? And why you increased it above .5 but did not decrease it below?

Response: Yes, in designing and writing the simulation study having separate pis for each group seemed to be the most convenient way of implementing a difference in pi between groups. On a more substantive level, the heterosexual group is used as the baseline, where no sibling effects seem to exist. It is thus reasonable to suppose that there are about as many older siblings as there are younger siblings in that group. We did not increase it below .5, because we wanted to show that the methods used by Blanchard et al. are unable to distinguish between an older sibling effect and a specific older brother effect. Tuning pi to be less than .5 would elicit a “younger sibling effect”, which we were not interested in and which is never discussed in the literature either.

Line 493: Same comment about Gelman and Stern as above.

Response: Please see our respective response (same as above).

Table 4, Model 9: The effect of interest is incorrectly written and copied from that of Model 8.

Response: We corrected this error.

Figure 1: Why does Model 9 not appear in this figure?

Response: Thank you for this further instance of attention to detail: done – corrected (i.e., added) in the revised ms.

C. Multiverse Meta-analysis

This section consists of two parts. The first part is a re-analysis of Blanchard (2018a, 2018b) and Blanchard et al. (2021), as well as the meta-analyses for the two sets consisting of all samples of men and women using the authors' preferred OR metric. It also presents the OBOR metric. The second part is a multiverse meta-analysis.

The first part seemed competently executed with one exception: it seems entirely unreasonable to use a fixed effects meta-analysis model for this data. The fixed effects model is almost never appropriate in biological and social science applications like this one, and the authors themselves note the variability within and across samples. Of course some point estimates and several lower interval estimates of heterogeneity for the random effects model are zero, but that is simply because the random effects model frequently has such zero estimates even when implausible (Chung, Rabe-Hesketh, and Choi, 2013; Chung, Rabe-Hesketh, Dorie, Gelman, and Liu, 2013). Please stick with the random effects model only.

Response: Thanks for the positive remarks. Concerning the fixed-effect model: It is a common misconception to discard the fixed-effect model as inappropriate in the light of effect size variability/heterogeneity. The fixed-effect model answers a different question than the random-effects model, in that it provides an average estimate across all of the available data. The random-effects model provides an estimate of the average effect of a population of effects sizes, which is assumed to follow a certain distribution (most often the normal distribution is assumed). Thus, the fixed-effect and random-effects summary estimates correspond to different models. The random-effects model also requires an estimate of the variance of this distribution. This information was already contained in the submitted ms. Please refer to the references cited in the ms., Rice et al. (2018) and Hedges and Vevea (1998), for further detail. The main argument is as follows: estimating the average effect underlying a set of observed studies is as legitimate a goal as is estimating the average effect of a population of effect sizes and researchers may choose that either (or both) of these models are relevant to their research question. So, our reporting of the fixed-effect models was not a decision motivated by the zero heterogeneity estimates (notice that we did not write something along these lines). Moreover, the goal of our meta-analyses is to map out the numerous possibilities and degrees of freedom, which are available and reasonable when analyzing the available data on the FBOE. One such way would be a fixed-effect model.

Further, I did not find the raindrop plots or albatross plots to be particularly useful. It would instead be useful to see more traditional forest plots with studies sorted in some reasonable order (e.g., effect size).

Response: The effect sizes are sorted with respect to their weight. This is the most reasonable way of sorting them in our opinion, as it shows that as the effects become heavier (regarding their weight in the meta-analysis), they also become smaller. As suggested, we have replaced the albatross plot by a funnel plot.

As for the second part (the multiverse meta-analysis), my concern is whether the ten "Which" factors specified on Line 822, etc. constitute true "researcher degrees of freedom" or whether the choices made by the researchers were indeed "good" or "right." For example, around Lines 700, the authors discuss the issue of whether gender dysphoric males should be classified as homosexual and allege this is "yet another unchecked researcher degree of freedom." Surely, however, it is a subject matter decision whether this is or is not a researcher degree of freedom: if most or all subject matter experts would generally agree that the vast majority or all of the gender dysphoric males are indeed homosexual in orientation, then it would seem not to be a researcher degree of freedom to code them as such. Indeed, on the contrary it would seem necessary to do so. Similarly with the choice of excluding bisexual and asexual groups as discussed in Line 693. Because I cannot well assess the choices made for these ten "Which" factors, I cannot place too much stock in the results of the multiverse meta-analysis.

Response: This is a very valuable and insightful remark to us, as we never thought about researcher degrees of freedom in terms of "good" or "bad". Researcher degrees of freedom can be based on subject matter insights – they are not questionable research practices (see Gelman & Loken, 2013, for a similar point). For each such decision which must be made there may be several reasonable alternatives, and each combination of alternatives may lead to different conclusions. This does not mean that researcher degrees of freedom are inherently bad, but it should raise awareness that they play an important part in the data analytic decisions - and most importantly, they are unavoidable. So, subject matter decisions such as classifying gender-dysphoric individuals as homosexual (which to us also appears most reasonable, although we are no subject matter experts) is also a decision which must be made and for which several alternatives are available (e.g., exclude gender-dysphoric individuals altogether, or categorize them into homo- and nonhomosexual, as Blanchard et al., who are subject-matter experts, did in some of their work).

Finally, I would again omit fixed effects models in this multiverse meta-analysis as they are not justified.

Response: Please refer to our above response.

Line 693-694: Why is this so obvious? Also, I think the tone here and on this whole page is bordering on unprofessional as it makes allegations which seem overly strong or at least not well founded.

Response: Thank you: this comment indeed was misplaced, being a remnant from an earlier ms. version. We removed it.

As for unfounded allegations: We assume that lines 691-729 (page 22 of the pdf) are referred and that this is based on our understanding of the term "researcher degrees of freedom". All statements on this page factually are correct (we provided references to support our statements). These are claims and decisions made by researchers investigating the FBOE. For instance, the meta-analysis by Blanchard (2018) explicitly stated that studies wherein sexual orientation was classified using a proxy were exempt from the meta-analysis, yet the meta-analysis contained samples which were categorized via a proxy variable. Although it is reasonable to classify gender-dysphoric individuals as homosexual, that classification nevertheless was made via a proxy variable. Moreover, inclusion criteria are, by definition, researcher degrees of freedom (decisions about the data-analytic plan).

Line 775: One can interpret the random effects model as having an assumed distribution as an inferential goal but one need not when one uses it. One can instead use it as a device for principled shrinkage estimation without believing anything about the population of studies. Therefore, the statement made here is too strong. I think it would be best to simply remove it.

Response: We were not precise enough in this text passage. The "inferential goal" referred to the model parameters that are estimated, which - in the random-effects model - consist of a mean and the variability of effect sizes. One need not interpret the variability (a practice which we criticize in this paragraph), but even using the random effects mean as just a shrinkage estimator requires an estimate (or the true value) of effect-size variability (tau; the equation of the random-effects mean estimator clearly requires a value for tau).

We clarified this statement.

Final comment:

The authors touch on an issue throughout that was not clear to me. In particular, Lines 293-297 Lines 721-723 Lines 1265-1275 all pertain to the unit of analysis. I have a question: in the original samples, are there cases where more than one sibling from a given family appears? If so, then treating siblings as independent seems problematic for obvious reasons. If however only one sibling per family appears, maybe this assumption does not really matter in practice. Can the authors discuss this issue at greater length? It seems like a pretty crucial one that merits more elaboration.

Response: Thanks for raising this point. We already discuss this at more length in our response to Reviewer 1. Bottom line: we were wrong in suggesting that this would be a problem. As for your specific questions, it is unknown whether there are cases of siblings from one family being in the same sample, but it appears that this is not the case across these studies. And yes, we no longer believe that the assumption matters, or even that it is not met.

Reviewer 3: Ruben Arslan

Basic reporting

The authors follow aspirational current reporting standards. Statistics in the manuscript are reported clearly, and the online supplement on OSF provides code, codebook, and data, where possible. The other criteria listed on PeerJ, which I will not repeat here, are met as well.

Experimental design

The approach is non-experimental, but I will use this space instead to comment that the authors identify a serious causal identification problem in the existing literature and solve it adequately. The research question (theoretical estimand) is clearly defined, there is a substantial knowledge gap (because past empirical estimands did not correctly capture the quantity of interest) and the study finally brings both in line. It's rigorously conducted and as far as I can tell does not lag the state of the art in meta-analysis.

Validity of the findings

Where possible, the underlying data has been provided (Tran et al.). The authors revisit a messy tranche of the literature and extract the signal from the noise. They are careful about limiting their conclusions to what the data can show and provide a wonderful counterbalance to past overclaims.

Additional Comments

This is extremely important work and it has evidently been a lot of work. The fraternal birth order effect has been canonized as a stylized fact, but the literature is apparently dominated by a confused approach to causal inference, wishful thinking, publication and selection bias, and a lack of openness and transparency. The evidence is much less clear than I would have thought, even though I did have my doubts. The authors bring a lot of light into the darkness here and have made a heroic effort to organize a seemingly very disordered literature. However, they struggle to summarize all their work effectively. My comments are intended to help that the implications are not lost on anyone and to make this work shine.

Response: Thank you for these set of positive remarks; we appreciate these very much.

---

1. Starting with the Abstract, I would ask the authors to make the paper punchier by using clear causal language and summarizing the main conclusions (the General Discussion does this better). So, the OBOR is inappropriate to isolate the causal effect of older brothers on male homosexuality from the effect of older sisters/older siblings in general. Is the empirical data

consistent with an effect of older siblings instead? Is the effect not specific to male homosexuality? Please make clear statements about this in the abstract.

Response: We have written the Abstract and Introduction sections of the manuscript almost entirely anew. Hopefully, the arguments are now presented more clearly. The reason we were (and still are) reluctant to use causal language is that we do not see how we make causal inferences. We assume the causal model (which we now describe in more detail) put forth by Blanchard et al. and assess technical claims about how to best investigate the single most important necessary observation required for this model to hold, which is homosexual men having more older brothers (relative to older sisters) than heterosexual men. We see, however, that this effect represents the "causal" or "theoretical" estimand.

Reviewer: I can understand that you are reluctant to draw strong causal inferences on the basis of the data yourself, but I think you have now succeeded at clarifying that Blanchard's is a causal claim.

The subphrase "almost no variation in the number of older brothers in men is attributable to sexual orientation" is very confusing. Obviously, being gay cannot cause anyone to have more older brothers, but the word "attributable" seems to prompt that interpretation. Shouldn't you say "the number of older brothers does not raise the odds of being homosexual in men"? And if it is still correlated with these odds (because it indirectly reflects confounding factors), you should say so too.

Response: We deliberately chose the word "attributable", as it is often used in the statistical literature and does not lent itself to causal connotations. Loosely, "attributable" is used to describe how much of the variation in one variable can be attributed to another variable in a statistical model. So this is what we were referring to: the association between the two variables is so small that neither explains the other one very well. We have changed our wording accordingly.

We are hesitant to conclude that the number of older brothers does not raise the odds of being homosexual in men, based on a fair assessment of our own results, because that would mean we would have to make a lot of assumptions about the underlying causal mechanisms, or would have to claim that the association between homosexual orientation and the number of older brothers is exactly zero (which we know is not true a priori). It is highly implausible that the association between number of older brothers and homosexual orientation is in fact 0. But what we can say - and we hope this is clearer now - is that, statistically speaking, any effect is so small, variable, and unspecific (consistent with a model wherein the statistical effect exists for men and women) that it may not be specific enough to draw substantive inferences such as "older brothers increase the odds of homosexual orientation via the MIH".

Edit: After having finished this paper for a second time, I return to the abstract. I think my high-level summary would be: "when analyzed correctly, the fraternal birth order effects is small, not specific to male homosexuality, and the existing evidence appears to be exaggerated by selection and publication bias."

Response: Yes, this is precisely what we tried to convey. Thank you, we have adopted the gist of this sentence in the revised Abstract.

Reviewer: Thank you. I like the revised Abstract. I understand the problem that we lack completely non-causal terms to describe associations. To my mind, only associated with and correlated with are really clear, but e.g. it's definitely depends on context (e.g. a correlation is an "effect size", even though it need to denote causal effect). As long as "X attributable to Y" has a different meaning than "Y attributable to X", I think there is a latent implication. Anyway, what you have now is pleasantly clear.

Throughout, I don't particularly like the phrase "an excess of older brothers". "excess" has a negative connotation, but more importantly the comparator is left unclear (compared to homosexual women/heterosexual men/women and heterosexual men?). I think almost everywhere this is used, you could more briefly say "have more older brothers than heterosexual men". I know Blanchard started it, but perhaps this is another bad practice of his, which verbally encodes the lack of clarity that his estimator shares. So, you need not repeat it.

Response: Thanks for pointing this out to us. We agree about possible negative connotations of the (statistical) term "excess". Accordingly, we have changed all occurrences of "excess of older brothers" to either "more older brothers relative to older sisters" ("relative" was suggested by Reviewer 1 and elegantly describes the group difference that Blanchard et al. are after) or "association between the number of older brothers and homosexual orientation" in our revised ms.

L50-55: Clear causal language would help make this section less confusing, but the substantive point is also wrong. If e.g. paternal age decreased the odds of homosexuality and fraternal birth order increased it, we might observe no correlational excess of older brothers among homosexual men, i.e. it'd be masked. I suggest you write something like "a causal interpretation necessitates that relevant confounding third variables have been controlled".

Response: Thanks for pointing this out to us. Reviewer 1 made a similar remark (see the respective response above). We amended our way of communicating that our perspective is conditional. We assume that the claims/model made by Blanchard et al. as given throughout the manuscript. In this model, no other variables/effects are relevant.

---

4. L56-60: I'd suggest to also give an expected rate (percentage) of homosexual children for no older brothers/1 older brother, as I find most people (including me) cannot naturally interpret percentage changes in odds.

Response: Done - we added (in the Introduction section) some examples of how the probability of homosexuality would change (on average) per additional older brother according to Blanchard et al.

5. L70-76: This is extremely confusing reasoning, but that is mainly on Blanchard et al. You are trying to say: this wasn't originally predicted, so the theory was retrofitted to explain it. You could and should be this frank, as it's a ridiculous auxiliary.

Response: Yes, it seems to be the case that this prediction was retrofitted. We rewrote these lines in the ms. text entirely.

6. L100-101 If these meta-analyses don't report an effect size, I think they are reviews at best.

Response: We agree, though Blanchard et al. referred to them as meta-analyses. Moreover, these "meta-analyses" report statistical tests incorporating an entire body of studies (via a sign test, for instance), which is not a feature common to reviews and more indicative of "some" way of quantitative research synthesis (although not a meta-analysis proper).

7. L109-110: A specific instance where any awake reader will ask themselves "an excess compared to whom?"

Response: Thanks for spotting this: changed accordingly.

8. L113 Starting with "Suppose". It'd pack more punch if you wrote simply: In the following, we show converging evidence that incorrect statistical reasoning has exaggerated the FBOE, and that, when corrected, the evidence is less certain and less specific to homosexual males than has been claimed. Together, these strands of evidence call the FBOE and the MIH into doubt.

Response: Agreed and appropriate changes made in ms. revision.

9. L144. I wonder whether what follows could be more easily digested, if you started with the proper, intuitive way to analyze (fraternal) birth order effects e.g. Rohrer et al. 2015. Then all the deviations from the standard way (individual as unit of analysis, regression framework) are more apparent.

Response: That's a good idea. We added a sentence highlighting other more intuitive ways of analyzing birth-order effects.

10. L459 Tran et al. Tran et al. (2019) repeated

Response: Corrected.

11. “We refer to this frequency as the false- positive rate. This definition of a false positive rate is conceptually different from what is understood as the false positive rate“. Why not use the established term “bias“ then?

We rewrote the description of the simulation study entirely, thus clarifying the point we wanted to make.

---

12. This probably varies from reader to reader, but I would have found only section 2, the forward simulation of data plus application of different “indices” easier to understand. Of course, it’s nice to prove the result using probability calculus (section 1), but know your audience. They were misled by the OBOR, so they might not be able to follow all this, and it could still feature in the supplement. As it stands, these sections accomplish the same goal, and though simulation and calculus aren’t redundant, you might lose some readers due to the sheer length of the combined sections. Which would be a shame, as the proper evidence synthesis in the MA is also really important.

Response: Yes, this is something Reviewer 1 also suggested. Reviewer 2, on the other hand, suggested removing the simulation altogether. For the sake of completeness of our presentation, we opted to keep it.

---

13. The meta-analysis seems to be well-done, although I’m not an expert on MA. I enjoyed reading the somewhat narrative style, in which critiques of previous attempts at evidence synthesis are melded with the explanation why this MA is conducted in a multiverse-style. That Blanchard and colleagues excluded a dataset with homosexual marriages but included phallometric testing of two-year-olds is just crazy. I guess I don’t have to wonder which method yielded the desired conclusion. I would appreciate seeing a standard funnel plot, as the Albatross plot is new to me and I don’t know how to read it to detect publication bias.

Response: We moved the funnel plot from the supplementary materials to the main section and the albatross plot from the main section to the supplementary materials.

---

15. Re the citation of Silberzahn et al. 2018, these authors (https://journals.sagepub.com/doi/abs/10.1177/23780231211024421) argue that the main problem was that the theoretical estimand was left unclear. This, at least, does not seem to be the case here (if only because the authors applied forensic analysis to Blanchard’s work).

Response: Thanks for pointing us to this excellent paper, which we now refer to in addition.

---

16. L1113: The results of Part III do not lend themselves to the confident

Response: Thanks!

---

17. L1124-1180: I am sorry to say this, but this entire section strikes me as entirely fallacious. The following paragraph makes sense until the "Nevertheless" "It is obvious that such factors have a different causal status in their relationship to the human sex ratio than does homosexual orientation of younger brothers, as the latter certainly cannot cause the sex ratio of older siblings to deviate from the population average. Nevertheless, one may use the extent of such observed variations in the sex ratio, as documented in various research sources (Gelman & Carlin, 2014; Gelman & Weakliem, 2009), and then determine where the variation in the sex ratio from the current results would rank among these and make an informed judgment about how plausible such results are. " The reason Gelman and colleagues sometimes poke fun at Kanazawa for reporting that nurses have many more daughters etc. is that the human sex ratio is very stable (for evolutionary reasons, though see Zietsch et al., 2020) and even extreme scenarios cause only little deviation. However, the fact that causal effects *on* HSR are known to be small does not mean that causal effects *of* HSR have to be small. If you want to make an evolutionary argument about why there shouldn't be large effects that cause homosexuality, you have to open a-whole-nother can of worms and cite an entirely different literature. If you go back to my point 1 about "variation in older brothers attributable to sexual orientation" this going back and forth when the causal direction is clearly only one way is very confused and confusing. Please strike the entire section and rest assured that your argument that the evidence is fickle is still very strong.

Response: This is a very valuable remark, and we agree. There is no reason to believe that a small causal effect of A on B implies that the effect of B on C is also small. It did not occur to us that the comparisons in this section could be interpreted causally. Our intention was to use the known statistical effects (non-causal interpretation of effects as comparisons between groups or strengths of associations, etc.) as a scale for comparing the statistical effects we observed in our analyses. We did not put forth any evolutionary arguments or substantive conclusions. We simply compared the magnitude of the statistical effects we obtained to the magnitude of effects usually encountered in the sex ratio literature. That being said, we now see that this was not optimal and that the comparisons do not make much sense, as all known effects from the literature actually are causal effects (or at least it is difficult to argue otherwise).

Hence, we removed this section accordingly, as per suggestion.

---

18. Given that you have done all this work to prepare the data for meta-analysis, I would appreciate seeing a meta-analysis of the effect of older siblings and older sisters in addition to the ratio. We've established that the FBOE/MIH approach has misled many for years, so let's take a step in the right direction.

Response: We think that generally this is an excellent idea, but we do not see how one could provide a such a meta-analysis. What would be the effect size? The odds ratio/difference in proportion of older versus younger siblings between homosexual and heterosexual individuals? A more adequate model would also look at both effects the older brother and older sibling effect simultaneously. However, these effects surely are correlated and hence modelling this association would require that we knew the correlation of each older brother-older sibling effect pair. Unfortunately, this type of information unavailable.

---

19. In the general discussion, I would appreciate seeing a recommendation that less explicitly instructs researchers on how to further investigate the FBOE. Instead, we should investigate the causal antecedents of homosexuality in an open-minded and rigorous way. One aspect of this would be to consider additional confounding factors such as parental age and parental loss. I would like this to be highlighted, as these confounds clearly should be controlled and acknowledging them should drive the final nail in the coffin of the idea that we can figure this out just with summary statistics at the sibship level. Frisch and Hviid (2006) seems to go in the right direction (large population sample, clear definition of homosexuality), but is not clear about the theoretical estimand and although many variables are adjusted for, some of them may be considered mediators (e.g., length of parental marriage is adjusted for even though it is influenced by parental loss and age).

Response: Thank you for this remark. Taking into consideration all of the reviewers' comments, we agree that our recommendations may well have been prescriptive. We rewrote this part of the General Discussion section, now emphasizing the suite of possible explanatory variables that in principle would need to be adjusted for.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed). Wiley.

Kronmal, R. A. (1993). Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, *156*(3), 379. https://doi.org/10.2307/2983064